Vol. 65, No. 3 May, 1958

Psychological Review

THEODORE M. NEWCOMB, Editor
University of Michigan

CONTENTS

Action Decrement and its Relation to Learning...EDWARD L. WALKER 129

Nonconstructural Aspects of Psychological Constructs.....W. W. MEISSNER 143

Elements of a Theory of Human Problem Solving.....ALLEN NEWELL, J. C. SHAW, AND HERBERT A. SIMON 151

Avoidance Conditioning Through Irradiation: A Note on Physiological Mechanisms and Psychological Implications.....JACK ARBIT 167

The Problem of Reductionism in Psychology.....RICHARD JESSOR 170

The Status of Inhibition in Hull's System: A Theoretical Revision.....H. GWYNNE JONES 179

PUBLISHED BIMONTHLY BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Consulting Editors

SOLOMON E. ASCH
ROBERT R. BLAKE
STUART W. COOK
CLYDE H. COOMBS
WILLIAM K. ESTES
LEON FESTINGER
W. R. GARNER
JAMES J. GIBSON
D. O. HEBB
HARRY HELSON
E. R. HILGARD
CARL I. HOVLAND
E. LOWELL KELLY
DAVID KRECH
ROBERT W. LEEPER
KENNETH MACCORQUODALE
ROBERT B. MACLEOD
DAVID C. MCCLELLAND
GEORGE A. MILLER
FREDERICK MOSTELLER
GARDNER MURPHY
OSCAR OKSER
CARL PFAFFMANN
CARROLL C. PRATT
JOHN P. SEWARD
DAVID SHAKOW
RICHARD L. SOLOMON
ELIOT STELLAR
S. S. STEVENS
ERIC TRIST
EDWARD L. WALKER
ROBERT W. WHITE

ARTHUR C. HOFFMAN, Managing Editor
HELEN ORR, Circulation Manager
HERBERT NELL, Editorial Assistant

The *Psychological Review* is devoted to theoretical articles of significance to any area of psychology. Except for occasional articles solicited by the Editor, manuscripts exceeding twelve printed pages (about 7,500 words) are not accepted. Ordinarily manuscripts which consist primarily of original reports of research should be submitted to other journals.

Because of the large number of manuscripts submitted, there is an inevitable publication lag of several months. Authors may avoid this delay if they are prepared to pay the costs of publishing their own articles; the appearance of articles by other contributors is not thereby delayed.

Manuscripts submitted for *regular publication* should now be sent to the *Editor-elect*, Richard L. Solomon, Emerson Hall, Harvard University, Cambridge 38, Massachusetts. Manuscripts submitted for *early* publication should be sent, until August 1, 1958, to the Editor, Theodore M. Newcomb, Doctoral Program in Social Psychology, University of Michigan, Ann Arbor, Michigan.

PUBLISHED BIMONTHLY BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
PRINCE AND LEMON STREETS, LANCASTER, PA.
1333 SIXTEENTH ST. N. W., WASHINGTON 6, D. C.
$8.00 volume $1.50 issue

Entered as second-class matter July 13, 1897, at the post office at Lancaster, Pa., under Act of Congress of March 3, 1879

Acceptance for mailing at the special rate of postage provided for in paragraph (d-2), Section 34.40, P. L. & R. of 1948, authorized Jan. 8, 1948

Send all communications, including address changes, to 1333 Sixteenth St. N.W., Washington 6, D. C. Address changes must arrive by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

© 1958 by the American Psychological Association, Inc.

VOL. 65, NO. 3 MAY, 1958

THE PSYCHOLOGICAL REVIEW

ACTION DECREMENT AND ITS RELATION TO LEARNING[1]

EDWARD L. WALKER

University of Michigan

It is the purpose of this paper to set forth some argument and some evidence that what is herein called *action decrement* and what has otherwise been labeled *reactive inhibition* (I_R) or *stimulus satiation* ($_sI$) is, at least in some instances, *positively* related to learning. This is in contrast to the usually held position that it is *negatively* related or antithetical to learning. Furthermore, it will be argued that when one tracks the course of the action decrement experimentally, one also, in fact, tracks the course of the perseverative consolidation which is frequently assumed to follow a period of training and is assumed to be necessary for subsequent performance.

The use of the term *action decrement* rather than any of the other available terms stems from the following beliefs. *Reactive inhibition* and *stimulus satiation* are no longer considered suitable labels because of the growing belief that any psychological action or event is followed by a temporary restriction upon its rearousal. It no longer seems certain that there is an exclusive relationship between action decrement and overt responses or to stimulus events, although it is admittedly difficult to deal with psychological events which are not stimulus-aroused. In short, the position taken is that the action decrement is a central event and the relevant symbol would thus be ($_{\epsilon}I$).

The contrasting points of view with respect to the role of action decrements can be elaborated on both a common-sense and on a theoretical level. On a common-sense level it is reasonable to view fatigue and fatigue-like states as biologically useful in preventing the organism from exceeding its resources and thus suffering damage. Such protection could be regarded as a sufficient function. Also on the common-sense level, we all know, or think we know, that a tired organism learns very inefficiently. Thus common sense tells us that Hull's (**5**, **6**, **7**) projection of these antithetical and detrimental effects to the theoretical level is eminently reasonable.

However, common sense sometimes provides contradictory propositions. It is possible to say that learning is not likely to occur without practice. Much practice is associated with much fatigue. The more one practices, the more one learns. Therefore, the more fatigue, the more learning. On a formal logical basis the conclusion is not certain, for there is a logical fallacy in the sequence of premises. Therefore, empirical demonstrations are required to establish the

[1] Most of the work reported in this paper has been made possible by a grant from the National Science Foundation.

truth of the conclusion. Several will be cited subsequently in this paper. Furthermore, the protective character of the fatigue-like state is retained, although the object of protection is a perseverative process rather than the whole organism.

The concept of perseverative consolidation in learning is an intriguing one. The concept in its most general form is quite simply stated. When an organism practices an action, there continues for a time a process which produces consolidation of the effects of the practice. If this consolidation process is interrupted or interfered with, by such events as a traumatic blow on the head or by practice on other similar material, as in the retroactive-inhibition paradigm, subsequent performance of the act will be deficient as a result of the interference. It is hypothesized that the action decrement is a direct result of the process which produces the perseverative consolidation. The biological utility of fatigue-like states, in this instance, lies in their capacity to protect the consolidation process from interference. A limited amount of practice, in any given instance, is efficacious. Beyond that limit fatigue is accompanied by a modification of the character of the process into some undesirable or inefficient form. Such modifications would constitute interference with the desired response. The fatigue-like state permits as much efficacious practice as the organism can tolerate without essentially modifying the character of the action system and its associated behavior.

Sharpless and Jasper (12) point out that there is a variety of decremental effects which occur as a function of continued or repeated stimulation. They make several distinctions. Among them there is a clear difference between "habituation" and "effector fatigue." For them,

> Habituation, like learning, is clearly referable to some form of plasticity in the central nervous system. . . . It does not depend on effector fatigue, since a response which has become habituated to a specific stimulus may still be elicited by an appropriate novel stimulus, whereas a fatigued effector will fail to respond, not only to the repeated stimulus, but to all other stimuli of the same relative intensity.

"Action decrement" is used in this paper to mean something very similar to "habituation" in Sharpless and Jasper. It is something like effector fatigue, which will accumulate in sufficient quantities to eventually modify the character of the response, usually in some undesirable form. Action decrement serves to delay or inhibit reactivation and thus to protect the consolidation process from the accumulation of effector fatigue to the point of serious modification of the character of the action system. Effector fatigue may be only one of several mechanisms of interference and may be chiefly associated with overt responses. Effector fatigue is not thought to be a significant factor in any of the situations considered in this paper.

The theory can be restated in a more concise, although not as yet in a very elegant form. *Any psychological action is followed by an action decrement—a lowered capacity for rearousal of the same event. The action decrement is a direct manifestation of the process of perseverative consolidation which is necessary for retention and subsequent performance. The action decrement persists for a limited time and then dissipates. Under many circumstances the dissipation of the action decrement is followed by an action increment which is learning or habit strength.*

The argument at this time is limited in the following sense. An attempt will be made to present positive, empirical evidence for the theory, but no attempt

will be made to account for every phenomenon in every situation in which concepts such as reactive inhibition and conditioned inhibition have been effectively and powerfully applied as negative factors, or factors antithetical to positive learning.

THE MODEL

What follows is a verbal model which has the potentiality of becoming, under intelligent modification, a rigorously deductive model, a neurophysiological model, a mathematical model, or all three at once. It is not clearly any of these in its present form.

Perhaps the best introduction is to state the problems inherent in what has already been said. It is fairly common practice to think of perseverative consolidation in terms of reverberatory circuits or systems. Reverberatory systems have two basic deficiencies, from the present point of view. There are grave suspicions that reverberatory activity with the initial stimulus as the sole impetus could not last as long as the consolidation process seems to last. Furthermore, one would expect such perseveration to present a condition in which it is *easier* to rearouse the action, rather than more difficult. That is, if some fraction of an action system continues in a state of reverberation, the threshold for that action should be lowered. It has been hypothesized here that the consolidation process is indexed by a decrement, a raised threshold.

Neurophysiological literature contains many empirical examples of the aftereffect of stimulation in systems more complex than a single fiber. One example which seems to offer promise may be seen in Fig. 1.

This figure is redrawn from *The Neurophysiological Basis of Mind* (1) by J. C. Eccles. The data in Eccles are taken from a study reported by Eccles and Rall in 1951 (2). A monosynaptic

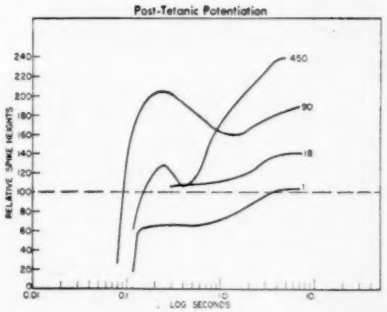

FIG. 1. Changes in sensitivity as a function of time after stimulation in a monosynaptic reflex system. (Redrawn on a log scale from Eccles and Rall, 1, 2.) The numbers 1, 18, 90, and 450 refer to the number of conditioning pulses delivered to nerve to the Flexor Longus Digitorum at a rate of 300 per second. Ordinate values represent height of test spike relative to first spike from conditioning stimulus.

reflex system is prepared in such a manner that the nerve trunk afferent to the reflex can be stimulated, and recordings of potentials can be taken on the efferent side of the synapse. The spike height of the first impulse is taken as the baseline (100%). The effects of any given number of conditioning pulses can then be traced by sending in a test pulse after a controlled delay. The height of the test spike is then expressed as a proportion of the first spike produced by the conditioning pulse. In the curve labeled "1" it can be seen that the second pulse was less than the first, if the test pulse was delivered any time during the first three seconds after the conditioning pulse. Thus there was an output decrement traceable for three seconds or more. If the conditioning stimulus involved 18 pulses, in this case delivered at a rate of about 300 per second, no decrement is visible in the graph; but there is a period of postsynaptic potentiation lasting for at least eight seconds. In the curves for 90 and 450 condition-

ing pulses, very brief decrements are plotted, followed by increments of 150% or more.

There are many reasons why it is dangerous to make much use of these data: these are motor neurones at the spinal cord level, the conditioning stimuli are not physiological stimuli, they amount to punishment of the fiber in a physoiological sense, etc. These data are offered only to support the reasonableness of the following propositions.

1. It is possible that the activation of neural tissue results in a decrement in output and sensitivity followed by an increment.

2. It is possible that the more intense the activity, the quicker the recovery, the greater the rate of recovery, and the greater the increment or potentiation.

These descriptive statements—of successive states of the neural tissue involved in a particular action and of the results of variation in intensity—then form the basis of a model.

When an animal alternates choices between two trials in a T maze, the phenomenon can be attributed to a decrement in the action system involved uniquely in the first choice. The raised threshold for this action forces the choice of the other alternative on the second trial. This situation can be used to illustrate and develop the model further. An animal is placed in the maze, turns left at the choice point, and is then removed. Only a small portion of the neural tissue available is assumed to have been involved in the action of choice, and it seems likely that many of the units so involved are fired several times in the process. It is possible to conceive that these units go into a decremental phase together, and in a brief period, possibly a second or two, recover more or less together, and then rise into a potentiated or supersensitive phase together. This picture would fall somewhere between 1 and 18 on the Eccles and Rall graph.

At this point let us invoke the Magoun Ascending Reticular System and its nonspecific cortical input. It is possible to conceive of either a rain of nonspecific impulses or of rhythmic waves of arousal activity at EEG frequencies as bringing about a refiring or recycling of the system after it rises into its supersensitive phase. It is obvious that such a mechanism could involve only a fraction of the original action system, for the animal does not repeatedly turn left at intervals from a fraction of a second to one or two seconds.

Assuming that only a portion, but a necessary portion of the system, goes through the process, the picture might be something like that seen in Fig. 2A, in

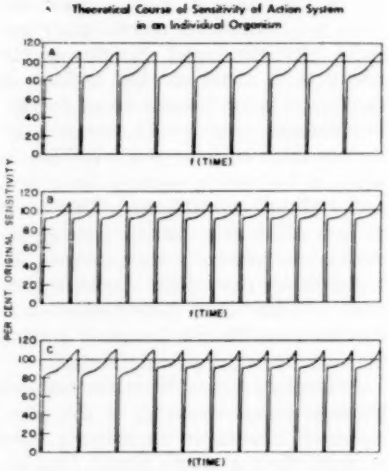

FIG. 2A. Theoretical recycling pattern for an action system driven by nonspecific input originating from a low drive state.

FIG. 2B. Theoretical recycling pattern under a higher drive state.

FIG. 2C. Theoretical recycling pattern under a low drive state with reward introduced after the third cycle. See text for explanation.

which the choice response is assumed to have occurred at the left. The fraction of the system goes into a decremental phase, and recovers into a potentiated phase. When it reaches an appropriate level, nonspecific input originating in some part of the reticular system can become effective in rearousing a fraction of the original action system, which then goes into a second decremental phase. This iterative process might well recur many times.

A further step involves the effects of different levels of motivation. Suppose one were to assume, as does Hebb (4), that differences in motivational level correspond in some way to differences in nonspecific output from the reticular system. According to the descriptive statement relative to the effects of intensity drawn from Eccles and Rall, one would expect an increase in the level of nonspecific input from an increase in motivational level and thus an increase in the intensity of the original action or any subsequent cycle of it. This would produce a quicker recovery in absolute time, a steeper curve of recovery, and a greater potential increment. Figure 2B illustrates the effect of a higher level of motivation on the shape of the cycle and the rate of iteration.

A final step involves an assignment of a role for reward within the model. Consideration leads to the conjecture that any reward condition might produce an increase in the level of nonspecific input, very much as if the level of motivation had been raised to a higher value. This condition is pictured, theoretically, in Fig. 2C, where reward is assumed to have been introduced after the third cycle. This effect of reward would fit the results of brain stimulation studies as reported by Olds and Milner (11), and as summarized by Olds (10). An implanted electrode with its tip in various parts of the brain can produce approximately the same effect on behavior as does water reward to thirsty animals. Both are conceptualized here as producing increases in the level on nonspecific input.

There are at least two factors which might cause this recycling process to cease. The integrity of the action system is dependent upon the condition that all units are originally fired at approximately the same time. At each recycling, a certain number of units could be expected to be lost because they did not recover in time to participate in the next recycling. Thus the size of the action system would grow progressively smaller through a process of fractionation. Furthermore, such a system has to some degree an afferent and an efferent end. One might conceive of the recycling process as being progressively restricted to the more efferent end. Both factors could be expected to produce a progressive decrease in the correlation between the excitability cycles of the units of the action system. The system would be conceived to have lost its dynamic integrity, and no further consolidation would occur from the moment when the correlation between the units reaches zero. When a unit drops out of the system, it can be expected to recover rather quickly to a semipermanent state of slightly lowered threshold, and when all perseverative activity has stopped, the combined threshold changes will constitute an increment in habit strength.

Certain characteristics of the model should probably be noted at this point. It describes a hypothetical pattern of behavior for a population of neural units which involve time units of the order of seconds. It may eventually be possible to derive this model from one which describes the behavior of individual units in time units of the order of milliseconds. Second, it will be noted that the burden of perseveration

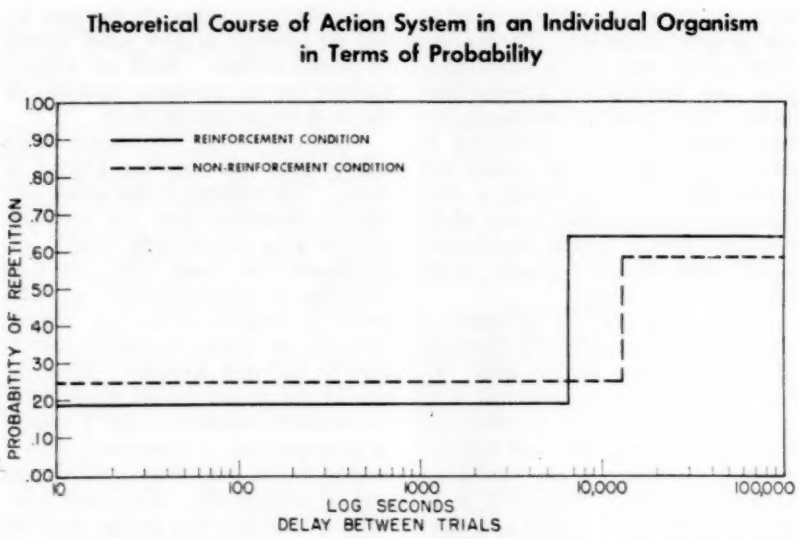

FIG. 3. Calculated probabilities of repetition for a nonreinforcement condition (from Fig. 2A), and from a reinforcement condition (from the latter part of Fig. 2C).

is in large part carried by the states of the units involved, where the state is a function of the original condition and the effect of the action. Thus, while perseverative recycling might be expected to be wiped out by electroconvulsive shock, it might well survive cooling in a hibernating animal. Cooling might slow or even stop the recycling process, but subsequent warming might bring it back into a perseverative condition. Finally, the model at this point is largely deterministic. At the next most molar level it becomes a probabilistic model.

Probabilities become factors in the model in several ways. As an action system rises into a potentiated phase, the precise point of recycling is subject to more or less random variation. Figure 2 might well have been drawn with different degrees of potentiation before recycling. Had this been done, there would have been an irregularity in the heights reached by the individual cycles, and there would have been a resulting variation in the time consumed by each cycle. Much more important, however, is the problem of the state of the system at the moment when the organism makes a second choice. If the system is in decrement, then the organism should choose the previously unchosen alternative. If it is in an incremental or potentiated phase, the choice should be to repeat the original choice. Since the time intervals involved in a single cycle in Fig. 2 should be in the general range of a second, and since experimental situations employed so far do not permit time control with any precision at this time range, the state of the system on the occasion of a second choice should be essentially a random sample of the available conditions. The prediction of what an animal will do on the occasion of a second choice can only be expressed in terms of probabilities.

The probability of repetition should be quite close to the ratio of time in a decremental state to time in a potentiated state. In Fig. 2A this ratio is 150:50, yielding a probability (p) of repetition of .25. In Fig. 2B and in the latter part of Fig. 2C, the ratio is 163:37, yielding a p value of .185. Thus the initial p values plotted in Fig. 3 are derived directly from Fig. 2.

In Fig. 3 the assumption is made that as long as any part of the action system is in an active state, as portrayed in Fig. 2, the probability of repetition will remain at a fixed value. When the active phase comes to an end, this p value will rise to a level slightly higher than it was before the activation of the system occurred originally. Obviously, repeated activation will lead eventually to a p value of nearly 1.00.

It is interesting to note that Fig. 3 conforms to the hypothesis previously stated by Walker (14) that the *greater* the decrement, the quicker the recovery and the greater the eventual increment (learning). In Fig. 1 the quick recovery and the greater increment are associated with the *smaller* apparent decrement. Thus one of the three correlated propositions is inverted in the process of transition from the deterministic level to the probabilistic one.

It is tentatively assumed that the amount of the decrement, the time required for recovery from the active phase, and the height of the increment will all be proportionally related. Whatever influences one will have a proportional effect on the other two. This should not only be true of experimentally manipulated variables such as level of motivation and amount of reward, but should describe individual differences between organisms as well. That is, if, under the same experimental conditions, two organisms show different probabilities of repetition during the decremental phase, the one with the lower probability of repeating should

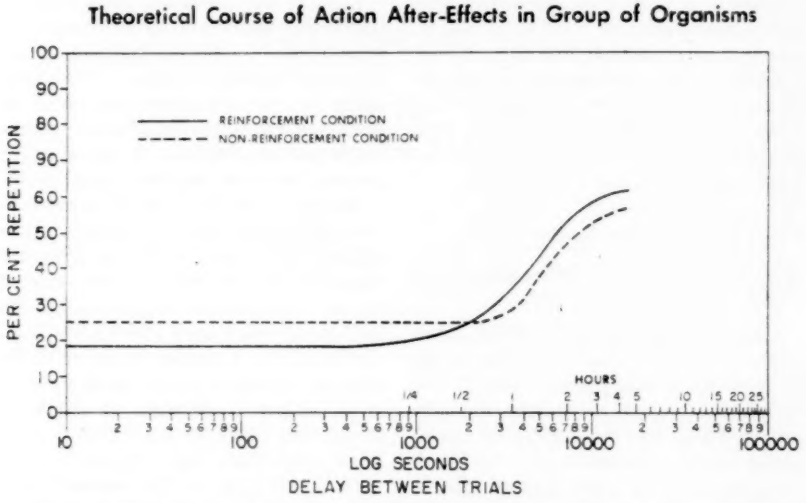

FIG. 4. Theoretical course of action aftereffects, expressed in terms of the percentage of a group of organisms which can be expected to repeat (or alternate) their first choice when given a second choice after a controlled delay.

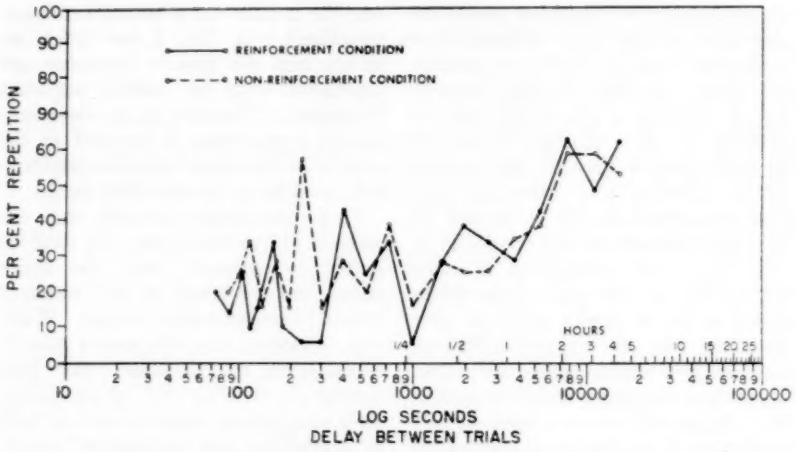

FIG. 5. Empirically determined course of action aftereffects, taken from Walker (13). Figure 4 is an approximate smooth fit to these data.

show quicker recovery and a greater increment of habit strength from a single trial. Since there is no immediately apparent way to assess the probability of repetition in a single organism, it is necessary to assume an essentially normal distribution of individual differences in probabilities; this permits calculation of the percentage of animals which can be expected to repeat at any given moment in time after a first choice of two alternatives. Figure 4 depicts this expectation.

EMPIRICAL EVIDENCE

The Duration of the Action Decrement and the Effect of Reward

The duration and course of the action decrement have been traced in an alteration study in a T maze (14). There were two groups of thirsty animals. One received water in both goal boxes; one group received no water. Each animal was given a single pair of choices on a given day, with the time between choices carefully controlled. Twenty-one different time intervals, ranging from about 80 seconds between trials to something over four hours, were investigated; there were 21 animals per point. The model, as specified, predicted two things.

First, it had previously been thought that action decrement, at least under its other names, would dissipate as an exponential function of time. The model suggested that the decrement would be sustained for a period without appreciable loss as a function of the recycling process, and then recovery would occur.

Second, when the reward and nonreward conditions were compared, it was expected that there would be a greater tendency to alternate in the reward group. This would be followed by a quicker recovery and then, after sufficient time, by a greater tendency to repeat, as a function of the habit strength acquired in a single trial. It should be noted that there is something a little nonsensical about this. A greater tendency to alternate in the rewarded animals would mean that if two thirsty animals turn left on the first trial, and if one is rewarded with water and the

other is not, the *rewarded* animal should have a greater tendency to turn the other way on the subsequent trial. But on the basis of acquired habit strength, alone, it should have a greater tendency to repeat its first choice at all times. The results may be seen in Fig. 5.

The expectation of a sustained decrement seems to be borne out by these results. There is a greater tendency to alternate (or lesser tendency to repeat) in the rewarded animals in the early stages. The quicker recovery for the rewarded group is suggested, but no more, and the habit strength from a single reward could not be demonstrated with the time and resources available. The theoretical curves in Fig. 4 are actually a smooth fit to the data in Fig. 5. The general form of the two curves and the relation between them were established prior to the experiment, but there was no way to suggest appropriate repetition percentages or time values.

These results seem to provide partial support for the theory and are certainly somewhat difficult to explain from a more traditional point of view.

Motivation and Alteration

The effort to deal with motivation has not been quite as successful as the effort to deal with reward. One previously unpublished study has been carried out using a shock grid throughout a T maze, with the goal boxes shock-free. A high level of shock (40 volts) was applied to the animal (his feet had been dipped in saline solution) throughout two trials. It was hoped to show a greater tendency to alternate than in the thirst and reward study, quicker recovery, and a greater increment in habit strength. There are only four points plotted in Fig. 6. The decrement is greater than in the previous study, but the recovery rate seems faster, since both of the middle points are out of the population of points in the thirst-reward study. The final point, at 24 hours, does not constitute a sig-

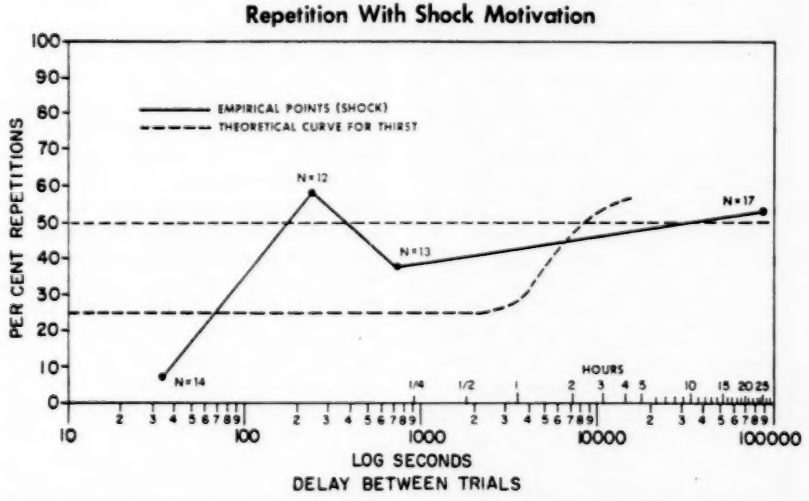

FIG. 6. The tendency to repeat choices under strong shock motivation with four different delays between trials, compared with the theoretical curve for strong thirst (see Fig. 4).

nificant demonstration of habit strength from a single trial.

Amount of Decrement and Ease of Learning

A very different kind of study emerges from a restatement of the theory in the following way. *Any factor capable of producing a big decrement should be capable of producing a big increment or rapid learning.*

The background for this experiment by Walker and Paradise (17) begins with a study by Glanzer (3), and a nearly identical one by Montgomery (9). Both reasoned that when an animal alternated in a T maze, one could not tell whether it was now turning right because it had previously turned left, or whether it now went to the white side because it had tried the black side the last time. Both built + shaped mazes that could be used as T mazes from either top or bottom by closing off one starting alley. If an animal is given a trial from one end and turns left to the black, and if his second trial is from the other starting alley, he can now alternate responses and turn right to the black, or repeat his left turn and alternate stimuli to go to the white, but he cannot alternate with respect to both stimulus and response. Both Glanser and Montgomery found alternation with respect to stimuli, and virtually none with respect to response.

Subsequently, Walker, Dember, Earl, and Karoly (15) went a step further, placing the + maze on wheels; with a somewhat more complex design they were able to offer animals a choice of alternation with respect to all three—stimulus, place, and response, or any one against the other two. It was found that stimulus was most important, and, by implication, was producing the greatest action decrement; place was next most important; and response was last.

Following that study, it was reasoned that the response could be made a significant factor in alternation if the feedback from the two responses was qualitatively very different. To accomplish this, a slightly modified + maze, called the side-winder or the contortion maze, was built. The animal was required to twist in three dimensions to enter the goal alley, and the responses to the two sides required two very different movement patterns. The previous design was now repeated by Walker, Dember, Earl, Fawl, and Karoly (16), and it was found that the response was now a considerable factor in alternation.

Since three factors could be combined or separated in the original (horizontal maze, H), and since there were four estimates of the differential capacity to produce alternation in the side-winder (vertical maze, V), there were available eight conditions which differed in their capacities to produce alternation. By inference, they differed in the relative depths of action decrement they produced. By hypothesis, they should differ in the same order in their efficacy as cues in learning. Therefore, eight learning conditions, four in each maze, were set up. In one condition in each maze, the reward remained with the same stimulus, the same place, and the same response, all at the same time (SPR-H and SPR-V). In another, the reward followed the stimulus, but was random with respect to response and place (S-H and S-V). In the third, the reward remained in a place, but was random with respect to the stimulus and response (P-H and P-V). In the fourth, the reward followed the response, but was random with respect to the stimulus and place (R-H and R-V). The results, with eight animals in each group, are shown in Fig. 7. The top figure shows the percentage of repetition (the reverse of alternation) found to each of the eight factors in the previous studies. Thus the letters SPR-V indicate that,

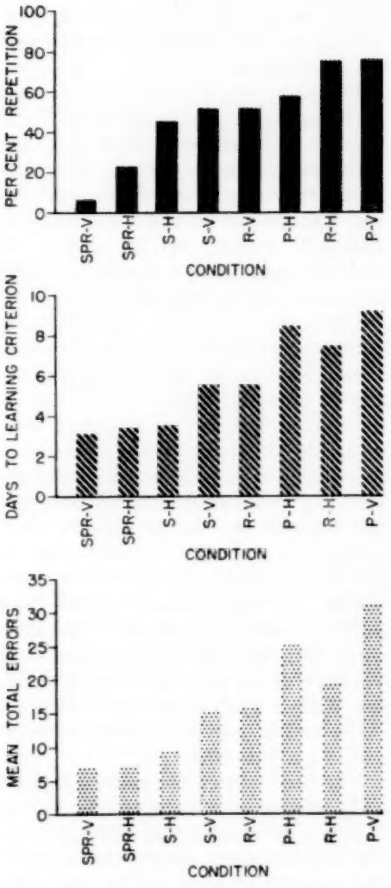

FIG. 7. Empirical test of proposition that the amount of decrement is positively related to rate of learning. Data taken from Walker and Paradise (16). See text for explanation of conditions.

when the animal was permitted to alternate with respect to all three in the vertical maze, there was very little tendency to repeat a choice but a great tendency to alternate, and thus a very large action decrement. On the other hand, the letters P-V mean that in the vertical maze, when the animal was offered an opportunity to alternate stimulus and response by repeating place, there was a strong tendency to repeat place, and, by inference, little action decrement.

The middle figure shows the number of days required to reach a criterion of learning in each of eight conditions. There are eight trials per day under thirst with water reward, and the criterion is 15 correct choices in 16 trials. The lower figure indicates the mean number of errors in each group. The fit is reasonably close and, to the extent that it is close, supports the theory.

Action Decrement, Stimulus Trace, and Perseverative Consolidation

It was asserted earlier that the process being tracked when one measures action decrement is perseverative consolidation. It also seems essentially true that a stimulus trace and a memory trace involve essentially the same mechanism. It was for this reason that the basic hypothesis of this paper was cast in terms of "any psychological action." This process consists of a progressive loss of members of the original population of neural units involved in the action.

According to the present hypothesis, the occurrence of the event in question, whether a presentation of a conditioned stimulus or a learning trial, serves to correlate the sensitivity cycles of a population of neural units. Driven by nonspecific input, a part of the system thus created goes through a neurally active, iterative, recycling process. At each cycle some of the original population of neural units is lost. The rate of loss will determine the duration of the active phase and the form of any measure of the effectiveness of the trace at any given point in time.

The form of such a trace is certainly a "simple decay function," as suggested by Hull (5) in 1943, and in certain instances is well fitted by a power func-

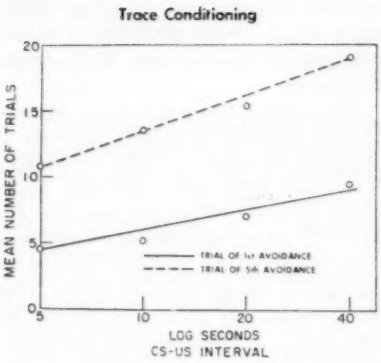

FIG. 8. Two measures of rate of learning of a trace-conditioned avoidance response with four different CS-US intervals. Data taken from Kamin (**8**), and drawn on a log abscissa.

tion, as indicated by Hull (**6**) in 1950. Kamin (**8**) has recently reported data on trace conditioning in dogs. He used CS-US intervals of 5, 10, 20, and 40 seconds, and reported a variety of measures of learning and extinction as a function of the duration of the trace. Certain of his measures of rate of learning seem to approximate a straight line when plotted on a log abscissa, as may be seen in Fig. 8. These particular measures were selected from several reported by Kamin, since it seemed likely that they were least affected by the fact that two animals in the 40-second group failed to learn to criterion.

What appear to be the least ambiguous results on the rate of decay of a memory trace or learning trace are those reported by Thompson and Dean (**13**). Following a learning trial, each recycling of the action system is conceived of as contribution to the habit strength acquired on that trial in proportion to the number of the original neural units remaining in the system in that cycle. If one should interrupt the recycling process at any given time after the learning trial, the amount of habit strength laid down by that trial should be describable as a linear function of log time between the trial and the interruption.

Unfortunately, it is extremely difficult to demonstrate the quantity of habit strength acquired on a single trial, and therefore even more difficult to demonstrate that a portion of that amount was lost through a given interruption. However, training can be massed so that one trial follows another in rapid succession and a large number of trials forcibly reactivate the action system without appreciable decay between trials. Under these conditions, a sufficient quantity of potential habit strength can be built up so that one would expect demonstrable differences in perseverative consolidation as a function of the interval between the end of training and the interruption. If the memory trace behaves as does the stimulus trace in the Kamin study (Fig. 8), then the amount of habit strength should be a linear increasing function of log time. Since the action decrement seemed to be virtually at an end after four hours in the Walker study (**14**), the theoretical curve in Fig. 9 is plotted to that point (see Figs. 4 and 5).

The study by Thompson and Dean (**13**) has the following character. They used a very simple maze with a starting box and a small **V** shaped choice chamber with two windows at the end, separated by a small board projecting out three inches from the window wall. They prehandled their animals for several days, and then, in preparing them for the critical learning period, taught them a simple grayness-discrimination problem for three days. They were motivated by shock applied if they stayed in the starting box more than 5 seconds or if they failed to choose a window after more than 30 seconds. There is no statement of how many shocks were required or through how much of the training period shocks

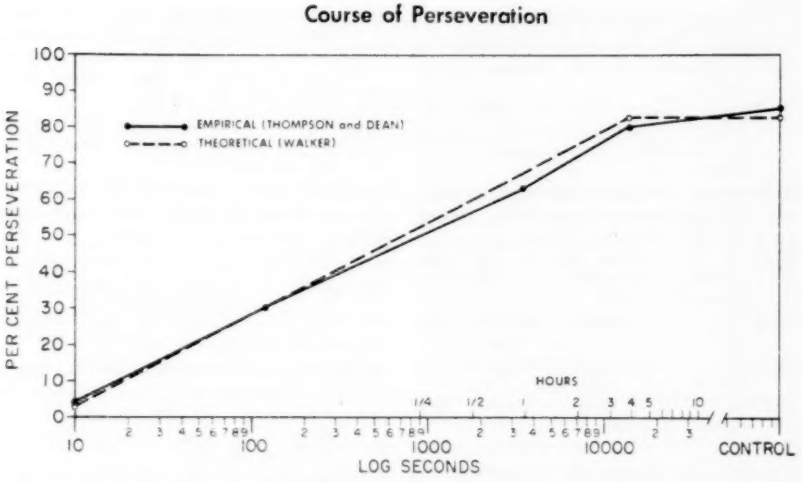

FIG. 9. A comparison of a theoretical prediction of the course of perseverating consolidation with empirical material on the effects of ECS on retention of a discrimination habit. Data taken from Thompson and Dean (12), and replotted here on a log abscissa.

were continued. Finally, Thompson and Dean gave their animals massed training on an easy horizontal-vertical problem. Each animal was trained on this problem in the shortest possible period of time; an average of about 28 minutes seems a likely estimate from their description. One group of such animals was given electroconvulsive shocks 10 seconds after the end of this training, another group was shocked two minutes after training, a third group one hour after training, a fourth group four hours after training, and a control group was not subjected to shocks. Forty-eight hours later they were retrained, and the relative savings in relearning were computed.

The results may be seen in Fig. 9. The dotted line is the theoretical prediction. The solid one is a replot of the error results reported by Thompson and Dean, now placed on a log scale. The apparent closeness is probably adventitious. From the theory, one would expect the duration of the action decrement to vary widely with the relative ease or difficulty of the problems presented, with the type and level of motive, with the magnitude and delay of reward, and with many other factors. The Walker study (14), on which the theoretical curve is predicated, and the study by Thompson and Dean (13) differed on all of these dimensions. Either the period of decrement and perseverative consolidation is much less subject to variation through the manipulation of these variables, or the closeness of fit in Fig. 9 is a most happy pair of combinations of variables.

In summary, it can be argued that there is some merit to the position that the action decrement represents a trace of the perseverative consolidation process, that it is thus positively related to learning, at least in some instances, and that one can expect operations which produce rapid learning to produce great decrements. Wherever it is applicable, this theory is in complete contradiction to the position that the fatigue-like decrement is detrimental to or antithetical to habit strength and learning.

REFERENCES

1. ECCLES, J. C. *The neurophysiological basis of mind.* Oxford: Clarendon Press, 1953.
2. ECCLES, J. C., & RALL, W. Effects induced in a monosynaptic reflex path by its activation. *J. Neurophysiol.*, 1951, **14**, 353–376.
3. GLANZER, M. The role of stimulus satiation in spontaneous alternation. *J. exp. Psychol.*, 1953, **45**, 387–393.
4. HEBB, D. O. Drives and the CNS (conceptual nervous system). *Psychol. Rev.*, 1955, **62**, 243–254.
5. HULL, C. L. *Principles of behavior.* New York: Appleton-Century, 1943.
6. HULL, C. L. *Essentials of behavior.* New Haven: Yale Univer. Press, 1950.
7. HULL, C. L. *A behavior system.* New Haven: Yale Univer. Press, 1952.
8. KAMIN, L. J. Traumatic avoidance learning: The effects of CS-US interval with a trace-conditioning procedure. *J. comp. physiol. Psychol.*, 1954, **47**, 65–72.
9. MONTGOMERY, K. C. Exploratory behavior and its relation to spontaneous alternation in a series of maze exposures. *J. comp. physiol. Psychol.*, 1952, **45**, 50–57.
10. OLDS, J. Physiological mechanisms of reward. In M. R. Jones (Ed.), *Nebraska symposium on motivation.* Lincoln: Univer. Nebraska Press, 1955. Pp. 73–138.
11. OLDS, J., & MILNER, P. Positive reinforcement produced by electrical stimulation in the septal area and other regions of the rat brain. *J. comp. physiol. Psychol.*, 1954, **47**, 419–427.
12. SHARPLESS, S., & JASPER, H. Habituation of the arousal reaction. *Brain*, 1956, **79**, 655–680.
13. THOMPSON, R., & DEAN, W. A further study on the retroactive effects of ECS. *J. comp. physiol. Psychol.*, 1955, **48**, 488–491.
14. WALKER, E. L. The course and duration of the reaction decrement and the influence of reward. *J. comp. physiol. Psychol.*, 1956, **49**, 167–176.
15. WALKER, E. L., DEMBER, W. N., EARL, R. W., & KAROLY, A. J. Choice alternation: I. Stimulus vs. place vs. response. *J. comp. physiol. Psychol.*, 1955, **48**, 19–23.
16. WALKER, E. L., DEMBER, W. N., EARL, R. W., FAWL, C. L., & KAROLY, A. J. Choice alternation: III. Response intensity vs. response discriminability. *J. comp. physiol. Psychol.*, 1955, **48**, 80–85.
17. WALKER, E. L., & PARADISE, N. E. An experimental test of the positive correlation between action decrement and learning. *J. exp. Psychol.*, in press.

(Received August 12, 1957)

Psychological Review
Vol. 65, No. 3, 1958

NONCONSTRUCTURAL ASPECTS OF PSYCHOLOGICAL CONSTRUCTS

W. W. MEISSNER

Saint Louis University

Much of what has been written on the subject of psychological theory construction, both on the part of psychological theorists and on the part of the philosophers of science, has been directed to the analysis and determination of the nature and function of the concepts which are found in such theoretical systems. The emphasis seems to have centered on the refinement of the methodology by which such concepts are formed, and on the meaning status of the constructs as they function in the system. Thus, operationism and the constructural method have received extensive treatment. They provide the method and the meaning for a scientific psychology.

Careful examination of the current psychological systems will reveal certain conceptual elements which function in the formulation and application of these theories, but which do not fall under the aegis of the formally defined and accepted methods by which the systems have been constructed. This paper will attempt to indicate some of these nonconstructural aspects, and briefly discuss one or two of the implications that flow from them.

Obviously, there are nonsystematic, nonscientific, nonconstructural operations and processes that enter into the construction of a theoretical system by way of necessary preliminaries. The scientist must read instruments, take measurements, feed rats, and a thousand other things, which are not directly incorporated into his theory, but which provide the data from which he will ultimately construct his system. We are not concerned with these pretheoretical elements. But we are concerned with other nonconstructural aspects which reveal themselves in the very construction and application of the theoretical constructs themselves, and may even alter the meaning of the theoretical concepts.

Our investigation will focus on three such nonconstructural elements: the attribution of real existence (as a real part of the organism) to theoretical constructs, the theoretical assumption of the construct as a quasi-real structure for purely methodological purposes (without asserting its real existence in the organism), and lastly, the experiential reinterpretation of theoretical concepts which have been previously defined formally by systematic methods.

Since the appearance of Bridgman's introduction of the concept of the operational definition into the methodology of psychology, psychological theorists have been deeply concerned over the adequacy of operationally defined constructs to function as descriptions of psychological phenomena (12). Bridgman's explicitation of the operational method focused attention on the character of the constructions which were related to any given complex of operations. Tolman (32, 33) identified and used the intervening variable as a type of construct mediating operationally defined variables. Although he seems to have intended the intervening variable as a purely functional and relational concept, it is not clear that this purely relational concept did not turn out to be something quite different (31).

The effort of MacCorquodale and Meehl (19) to clarify the terminology

applied to these constructions was a start in the direction of identifying the kinds of concepts which could function in a psychological theory. They distinguished the intervening variable from the hypothetical construct. The intervening variable was an abstract, empirical relation, which merely served to organize lower level empirical laws. The hypothetical construct, however, involved some supposition of entities or processes, which were not directly observed, and therefore were not included in the original definitions. Essentially, this "surplus meaning" was taken to be the implication of the existence of entities and the occurrence of events in the real subject of investigation. However, it seems that this interpretation of hypothetical constructs and their "surplus meaning" is rather limited and has led to a certain amount of unnecessary confusion.

Supposing for the moment that "surplus meaning" has reference primarily to the suggestion of existence (it may signify other things),then what interpretations can be put on this "existence"? In the original sense of real existence, as suggested by MacCorquodale and Meehl, there is necessary a further distinction. "Real existence" of an inferred entity can be taken as meaning that the inferred entity, as formally understood, really exists as part of the intrinsic, objective, noninferred, real structure, which constitutes the object of investigation. Or, the "real existence" of an inferred entity can be understood to mean only that the construct is formed and intended as a theoretical correlative of the real structure of the object. In this sense, a hypothetical construct is said to denote (10). Lindzey (18) has expressed the need for this sort of clarification in his attempt at the description of what he termed "conventional constructs." This type of construct involves an inference about the reality of the organism no less that any hypothetical construct—but it is precisely an inference about the organism *as it is known in the theory*. It is not an inference about the real structure of the organism itself. The former sense, if found in the formulation and interpretation of psychological theories, constitutes a gross misunderstanding of the nature of theoretical constructions, as well as a misapplication of methodology. The reasons for this will be discussed below. The latter sense, however, is a legitimate systematic concept, and is the only acceptable and possible meaning of "real existence" as applied to hypothetical constructs. Since this sort of "surplus meaning" is not found in the constructed definition of the concept, however, it is a nonconstructural element within the framework of the theory.

This theoretical-real existence must not be confused with the kind of existence which is applied to a construct by reason of its contextual consistency within the framework of its proper theory. This sort of "existence" is nothing but the coherence of logical relations among the constructural concepts which compose the theory. The logical coherence of constructs in a theory can not be taken as indicative of anything beyond the theory itself. This sort of *systemic existence* (1) is common to both intervening variables and hypothetical constructs. The "surplus meaning" involving "real existence" of hypothetical constructs must be reserved to theoretical-real existence in the latter sense of a theoretical correlate described above.

The failure to recognize this distinction between meanings of real existence has led to a reconfusion of terminology (25). The supposition of theoretical-real existence has sometimes been assimilated to the intervening variable (22, 23), with the result that the intervening variable is no longer merely

intervening, and the hypothetical construct is reduced to the level of a preliminary and imperfect stage in the development of a fully operational theory based on the intervening-variable type of construct alone. If Feigl's "existential assumption" can be interpreted in this sense of theoretical-real existence, it would seem that this plea for the return of the hypothetical construct is fundamentally an appeal to this sort of distinction in the meaning of "real existence" (7, 8, 9).

Correlative to the issue of real existence is the question of the conceptual status of constructs and constructural models as they function in theory. The issue of the legitimacy or illegitimacy of the flat assertion that theoretical entities have real existence will be decided in terms of our understanding of the nature of these concepts. There is no question of the legitimacy of constructural concepts, but only of the assertion of their existence. Generally speaking, methodologists are aware that a theoretical model can not be understood as representative of the real structure of some real entity (28, 31). This is fairly obvious in the case of the intervening variable, whose characteristics are of a purely formal nature. But in the case of the hypothetical construct, where some sort of objectivized (not objective) structure is implied, the issue is raised again and again.

A correct appraisal of the nature of constructural methodology has led many theorists to evaluate the legitimate "surplus meaning" (i.e., theoretical-real existence) of the hypothetical construct as an extrasystematic and, therefore, nonconstructural element, which introduces an impure fraction into the constructural purity of the theory. However, the surplus enters into the understanding of such a construct only to the extent that the construct itself is structural (15). The hypothetical construct differs from an intervening variable precisely in this, that it hypothesizes some sort of structure to explain the relation of lower-level variables (25). Obviously, a scientist can not think of structure, as opposed to a merely formal or mathematical function, without thinking of it in terms of "theoretical-real existence." This does not imply anything about the *real* structure of any *real* thing, but it does imply that the theorist is thinking of this construct as an analogue to the real structure of a real thing.

Theoretical models must be thought of then as research tools, which the theorist forms without direct dependence on or inference from the immediate data of experience. The constructural-theoretical model is related to the ultimate real structure of the thing under investigation through a complex set of relations, which involve certain specified methodological procedures, selective processes, and measurement techniques. The point here is that theoretical constructs are not directly tied down or determined by the real structure of their objects. If they were so tied to the structure of the thing, there would be little room for the multiple hypothetical models we find in psychology, nor could the theorist enjoy the relative freedom in structuring the model that is now his. It is for this reason that the assertion of the noninferred real existence of such theoretical entities must be looked upon as illegitimate.

Models should not really be thought of as primarily imitative, but rather as a sort of adaptable pattern which can be compared with the available data of experiment to find out what adaptations are to be made and in what direction the course of our investigation must be turned (21). Although it is not clear to me what the criteria of usefulness are to be in the development of a psychological theory, the recent emphasis on "usefulness" of theoretical constructs,

at least for the formulation of research hypotheses, marks an encouraging shift from the former preoccupation with reality (30). Even on logical grounds, it has been pointed out (20) that the common procedure of testing theories by inference and confirmation is open to the possible fallacy that the consequence is unique for that theory. If Consequence B is deduced from Theory A and experimentally confirmed, we have added to the probability of Theory A; but we cannot assume that the same Consequence B may not be implied by some alternative theory. The hypothetical construct is a surrogate or symbol which is constructed in the theorist's mind as a substitute for the real structure of the organism, and which is only indirectly related to the structure of that organism. The whole direction or impulse of theory construction is to the progressive formulation of better theoretical constructions which are capable of more rigorous evaluation and which will gradually converge on the real structure of the organism (24).

Feigl (6) has expressed the belief that higher-order hypothetical constructs, which are built on and constructed from operations, will gain real insight into the natures of things. But in a later writing (8), he recognizes that the meaning of a psychological concept is generally not identical with its defining operations. If we are able to speak of insight into the natures of things (whatever that may mean in the context of constructural theory), whatever meaning is contained in higher-level constructs must be reflected in the lower-level definitions in terms of observable data. If meaning and definition are not identical, we must explain the difference. I do not mean that we must explain the fact of a difference, but we must define the difference. Otherwise our procedure is unscientific and methodologically unsound.

Along this same line, Krech (17) has pointed out that hypothetical constructs could not function in a purely psychological psychology, since the purely psychological data are beyond the reach of scientific inquiry. Because the experience of a private datum is an unverifiable event, and because science is concerned only with verifiable events, some would claim that a scientific psychology of private experience is a contradiction (32). Either psychology proceeds behavioristically and is a science, or it treats of the private datum and relinquishes its claim to scientific status (4). Thus Krech has indicated that the concepts of tensions, needs, and cognitive structures are sometimes defined as hypothetical constructs, and sometimes as conscious experience or behavior (17). This means that there has been a shift in the formal meaning of these concepts, and we are dealing with two concepts instead of only one. One is a formally defined hypothetical construct, which functions as an integral part of a theory. As such, it is linked to the observable data, from which it is structured, by low-level empirical laws and operational definitions. The other concept, which represents the redefinition of the previous hypothetical construct in terms of experience, is more difficult to delineate because of the fact that it is rarely, if ever, given any sort of explicit or formal presentation. However, in general, it seems to take the form of a subtle reinterpretation of a previously defined concept in terms of some sort of phenomenological experience. This sort of shift in meaning has been pointed out as due to a failure to distinguish between the operational meaning of intervening variables or hypothetical constructs and the intuitive properties ascribed to them (14).

This shift in the interpretation of concepts must not be confused with the possible shift that sometimes occurs in the construction of theoretical systems

between intervening variables and hypothetical constructs. It has sometimes happened that a concept is defined and constructed as a purely functional intervening variable, but later in the development of the theory it acquires a denotative meaning and a certain amount of structural connotation, which qualifies it as a hypothetical construct (10). Both Tolman (26) and Hull (10, 19, 25, 27) have been accused of permitting this subtle reorientation in the meaning of their concepts. Very likely, Bergmann's dismissal of the distinction between intervening variable and hypothetical construct as a pseudodistinction (3) stems from insufficient attention to the possibility of this sort of subtle shift in the formal meaning of concepts.

As long as the surplus meaning, which constitutes the concept as a hypothetical construct, involves no more than the assumption of theoretical-real existence, the procedure is by no means illegitimate, and can be scientifically fruitful. It is this sense, I believe, that Feigl has spoken of "existential assumptions" (7, 8). However, if the surplus meaning involves the implication of real existence (involving attribution of the hypothetical construct to the organism as part of the real structure of the organism), the connotation is illegitimate and unscientific, as we have already pointed out. There seems to be a certain amount of evidence to suggest that Hull was not always able to avoid this error. In regard to the *sHr*, he wrote that "it exists as an organization as yet largely unknown, hidden within the complex structure of the nervous system" (11, p. 102). Likewise, MacCorquodale and Meehl were quick to suggest that Hull's early articles on conditioning were written because Hull "considered that the hypothetical events represented in his diagrams may have actually *occurred* and that the occurrence of the events represents the underlying truth about the learning phenomena he dealt with" (19, pp. 104–105).

In our discussion thus far, we have seen that this sort of implication of the real existence of an inferred entity as a part of the real structure of an organism is extrasystematic and to that extent, nonconstructural. This also constitutes an illegitimate inference, which is not only not supported by the defining method, but is contrary to the very nature of the scientific constructural methodology. We have also seen that the theoretical assumption of the constructed concept as correlative, and perhaps suggestive of the real structure of the organism, is again extrasystematic and nonconstructural, but in this case the assumption of existence (which is equivalently an assumption of *non*existence) is entirely legitimate, and in fact a necessary correlate of the development of theoretical concepts. We have further suggested a third kind of nonconstructural ingredient which seems to function in psychological theory: the reinterpretation of systematically constructed concepts in terms of experiential variables. This last nonconstructural aspect deserves further consideration.

Spence (31) has already pointed out that the method of field theorists like Lewin depends heavily on phenomenological introspection for the introduction of theoretical concepts. If these concepts are then defined systematically in terms of given observables, they function as integral parts of a scientifically constructed theory. But, if they continue to be applied and interpreted in terms of the original experiential context in which they arose, they constitute something else by way of formal concepts. It is quite possible that the correlation of empirical constructs with formal terms according to the logic of dynamics, as Koch (16) has analyzed them in Lewin's system, presents a dif-

ficulty precisely in so far as the empirical constructs are functioning at a phenomenological level and the formal terms are presented at a constructural level.

An example may help to clarify the problem. Hull (11) introduces his concept of habit strength as a functional stimulus-response relation between stimulus energy, response, nature and number of reinforcements, and the reaction potential, which is measured in terms of the actual response. This intervening-variable type of concept is then conceived as a dynamic relationship between the afferent and efferent neural impulses. This seems to be more of a hypothetical construct, but the definition of the term is still couched in systematic and operational terms, and the switch from intèrvening variable to hypothetical construct is still acceptable. However, at one point Hull tells us that "it would appear that the presence of internal conditions or neural organizations, such as habit structures (*sHr*) and reaction potentials (*sEr*) are also reportable" by introspection (13, p. 344). He is also explicit in stating his conception of receptor-effector connections in terms of common sense habits (11). If Hull is able to identify these constructs by introspective methods, it follows that the concept, as operationally defined and identified, is now being reinterpreted and reidentified in terms of experiential elements which are not included in the systematic formulation of these concepts. Phenomenological introspection is not restricted to field theorists alone, so it seems.

The introduction of these nonconstructural elements into even the most sophisticated of behavioral systems raises a serious question. Snygg (29) has raised the issue of a phenomenological psychology. But, like Tolman (32) before him, he draws the line between the phenomenological and the objective in such a way as to give us a none too happy choice. Are we forced to make such a choice? The scientific method, as it has been transferred from natural sciences to psychology, seems to have exhausted a good deal of its potentiality (5). Further, as a method, it is not equipped to deal with purely psychological data. As a result, and this is the point I have tried to make in this paper, psychological theory has resorted to a subtle, almost unconscious, translation of its systematic concepts in terms of psychological experience. This step has been necessary, in a sense, in order that psychological constructs assume some semblance of pertinence and meaningfulness for the human psychological situation. And this experiential interpolation is a necessary development in all psychological theory, for it is only in terms of this sort of interpolation that psychological theory can function as significant for the problems of applied human psychology, and for the prediction and control of behavior.

Rather than interpret this analysis as disjunctive and exclusive—and this seems to be the prevailing attitude of psychological theorists—it might prove promising to take it as conjunctive and inclusive. I do not mean that the experiential and objective aspects are reducible to common grounds. Nor do I mean that one or the other should be eliminated. The objective, scientific techniques developed by experimental psychology are invaluable, and any attempt at devaluating or reducing them would be unrealistic, to say the least. Experimental methods are a necessary means of introducing control, confirmation, and validation into theory construction. But it does not seem that these methods are in themselves sufficient to meet the demands of a scientific psychology. Our theories are now subject to experiential reinterpretation which takes on the color of a necessary

adjunct to their significance and applicability. As the situation now stands, this experiential interpolation is unhealthy from the standpoint of scientific theory because it is extrasystematic and has no formal place in our theory construction. For this reason, the more rigorous theorists have tended to reject it explicitly, although they cannot avoid it implicitly (2).

But, it is not entirely clear that phenomenological or experiential aspects need be so excluded. If these elements can be made explicit in our theories and can be justified in our methodologies, the present status of psychology would be much healthier than it is. The formulation of a distinctively psychological methodology, based on the inclusion of both objective and experiential methods, and giving rise to theoretical construction of a distinctively psychological character, is a very tempting goal. The reconstruction of such a distinctive and comprehensive methodology must be left to the future. But at least a first step in its direction must be taken by pointing out the revelant elements as they function in the theories we have at the present time.

SUMMARY

Three nonconstructural elements, which appear in current psychological constructural systems, are singled out and briefly discussed: (*a*) attribution to a theoretical construct of "real existence," as if the construct were actually a description of the real structure of the organism, rather than of the organism *as it is known in the theory;* (*b*) attribution of "theoretical-real existence" to a construct, without attributing the defined structure to the real organism which is merely a methodological assumption and implies nothing about the real organisms; (*c*) phenomenological or experiential reinterpretation of theoretical constructs. Of these nonconstructural aspects, the first is illegitimate, the second is legitimate and scientifically sound, and the last is discussed in the light of possible evolution of a distinctively psychological methodology which would incorporate both phenomenological and behavioristic elements.

REFERENCES

1. BECK, L. W. Constructions and inferred entities. *Phil. Sci.*, 1950, **17**, 74–86.
2. BERGMANN, G. On some methodological problems of psychology. *Phil. Sci.*, 1940, **7**, 205–219.
3. BERGMANN, G. Theoretical psychology. *Ann. Rev. Psychol.*, 1953, **4**, 435–458.
4. CONWAY, B. J. W. Philosophy, metaphysics and psychology. In *Proceedings of the Second International Congress of the International Union for the Philosophy of Science, 1954.* Vol. V. Zurich: Neuchatel Suisse, 1955. Pp. 79–80.
5. FARRELL, B. A. On the limits of experimental psychology. *Brit. J. Psychol.*, 1955, **46**, 165–177.
6. FEIGL, H. Rejoinders and second thoughts (Symposium on operationism). *Psychol. Rev.*, 1945, **52**, 284–288.
7. FEIGL, H. Existential hypotheses. *Phil. Sci.*, 1950, **17**, 35–62.
8. FEIGL, H. Principles and problems of theory construction in psychology. In W. Dennis (Ed.), *Current trends in psychological theory.* Pittsburgh: Univer. Pittsburgh Press, 1951. Pp. 179–213.
9. FEIGL, H. Functionalism, psychological theory, and the uniting sciences: Some discussion remarks. *Psychol. Rev.*, 1955, **62**, 232–235.
10. GINSBERG, A. Hypothetical constructs and intervening variables. *Psychol. Rev.*, 1954, **61**, 119–131.
11. HULL, C. L. *Principles of behavior.* New York: Appleton-Century, 1943.
12. HULL, C. L. The problem of intervening variables in molar behavior theory. *Psychol. Rev.*, 1943, **50**, 273–291.
13. HULL, C. L. *A behavior system.* New Haven: Yale Univer. Press, 1952.
14. KENDLER, H. H. "What is learned?"—A theoretical blind alley. *Psychol. Rev.*, 1952, **59**, 269–277.
15. KESSEN, W., & KIMBLE, G. A. "Dynamic systems" and theory construction. *Psychol. Rev.*, 1952, **59**, 263–267.

16. KOCH, S. The logical character of the motivation concept. *Psychol. Rev.*, 1941, **48**, 15–37; 127–154.
17. KRECH, D. Dynamic systems, psychological fields, and hypothetical constructs. *Psychol. Rev.*, 1950, **57**, 283–290.
18. LINDZEY, G. Hypothetical constructs, conventional constructs, and the use of physiological data in psychological theory. *Psychiatry*, 1953, **16**, 27–33.
19. MACCORQUODALE, K., & MEEHL, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.*, 1948, **55**, 95–107.
20. MCGUIGAN, F. J. Confirmation of theories in psychology. *Psychol. Rev.*, 1956, **63**, 98–103.
21. MACKAY, D. M. On comparing the brain with machines. *Amer. Scientist*, 1954, **42**, 261–268.
22. MARX, M. H. The general nature of theory construction. In M. H. Marx (Ed.), *Psychological theory*. New York: MacMillan, 1951. Pp. 4–19.
23. MARX, M. H. Intervening variable or hypothetical construct? *Psychol. Rev.*, 1951, **58**, 235–247.
24. MARZOCCO, F. N. The character of psychological theories. *J. gen. Psychol.*, 1955, **52**, 197–203.
25. MAZE, J. R. Do intervening variables intervene? *Psychol. Rev.*, 1954, **61**, 226–234.
26. O'NEIL, W. M. Hypothetical terms and relations in psychological theorising. *Brit. J. Psychol.*, 1953, **44**, 211–220.
27. ROZEBOOM, W. W. Mediation variables in scientific theory. *Psychol. Rev.*, 1956, **63**, 249–264.
28. SHOBEN, E. J. Psychological theory construction and the psychologist. *J. gen. Psychol.*, 1955, **52**, 181–188.
29. SNYGG, D. The need for a phenomenological system of psychology. *Psychol. Rev.*, 1941, **48**, 404–424.
30. SNYGG, D. Scientific method in psychology. *J. gen. Psychol.*, 1955, **52**, 189–196.
31. SPENCE, K. W. The nature of theory construction in contemporary psychology. *Psychol. Rev.*, 1944, **51**, 47–68.
32. TOLMAN, E. C. Psychology versus immediate experience. *Phil. Sci.*, 1935, **2**, 356–380.
33. TOLMAN, E. C. Operational behaviorism and current trends in psychology. In *Proc. 25th Ann. Celebr. Inaug. Grad. Stud.* Los Angeles: Univer. Southern California Press, 1936. Pp. 89–103.

(Received May 31, 1957)

Psychological Review
Vol. 65, No. 3, 1958

ELEMENTS OF A THEORY OF HUMAN PROBLEM SOLVING

ALLEN NEWELL, J. C. SHAW
The RAND Corporation
AND HERBERT A. SIMON
Carnegie Institute of Technology

In this paper we shall set forth the elements of a theory of human problem solving, together with some evidence for its validity drawn from the currently accepted facts about the nature of problem solving. What questions should a theory of problem solving answer? First, it should predict the performance of a problem solver handling specified tasks. It should explain how human problem solving takes place: what processes are used, and what mechanisms perform these processes. It should predict the incidental phenomena that accompany problem solving, and the relation of these to the problem-solving process. For example, it should account for "set" and for the apparent discontinuities that are sometimes called "insight." It should show how changes in the attendant conditions—both changes "inside" the problem solver and changes in the task confronting him—alter problem-solving behavior. It should explain how specific and general problem-solving skills are learned, and what it is that the problem solver "has" when he has learned them.

Information Processing Systems

Questions about problem-solving behavior can be answered at various levels and in varying degrees of detail. The theory to be described here explains problem-solving behavior in terms of what we shall call *information processes.* If one considers the organism to consist of effectors, receptors, and a control system for joining these, then this theory is mostly a theory of the control system. It avoids most questions of sensory and motor activities. The theory postulates:

1. A control system consisting of a number of *memories,* which contain symbolized information and are interconnected by various ordering relations. The theory is not at all concerned with the physical structures that allow this symbolization, nor with any properties of the memories and symbols other than those it explicitly states.

2. A number of *primitive information processes,* which operate on the information in the memories. Each primitive process is a perfectly definite operation for which known physical mechanisms exist. (The mechanisms are not necessarily known to exist in the human brain, however—we are only concerned that the processes be described without ambiguity.)

3. A perfectly definite set of rules for combining these processes into whole *programs* of processing. From a program it is possible to deduce unequivocally what externally observable behaviors will be generated.

At this level of theorizing, *an explanation of an observed behavior of the organism is provided by a program of primitive information processes that generates this behavior.*

A program viewed as a theory of behavior is highly specific: it describes one organism in a particular class of situations. When either the situation or the organism is changed, the program must be modified. The program can be used as a theory—that is, as a predictor of behavior—in two distinct

ways. First, it makes many precise predictions that can be tested in detail regarding the area of behavior it is designed to handle. For example, the theory considered in this paper predicts exactly how much difficulty an organism with the specified program will encounter in solving each of a series of mathematical problems: which of the problems it will solve, how much time (up to a proportionality constant) will be spent on each, and so on.

Second, there will be important qualitative similarities among the programs that an organism uses in various situations, and among the programs used by different organisms in a given situation. The program that a human subject uses to solve mathematical problems will be similar in many respects to the program he uses to choose a move in chess; the program one subject uses for any such task will resemble the programs used by other subjects possessing similar training and abilities. If there were no such similarities, if each subject and each task were completely idiosyncratic, there could be no theory of human problem solving. Moreover, there is some positive evidence, as we shall see, that such similarities and general characteristics of problem-solving processes do exist.

In this paper we shall limit ourselves to this second kind of validation of our theory of problem solving. We shall predict qualitative characteristics of human problem-solving behavior and compare them with those that have already been observed and described. Since all of the available data on the psychology of human problem solving are of this qualitative kind, no more detailed test of a program is possible at present. The more precise validation must wait upon new experimental work.[1]

In succeeding sections we shall describe an information-processing program for discovering proofs for theorems in logic. We shall compare its behavior qualitatively with that of human problem solvers. In general, the processes that compose the program are familiar from everyday experience and from research on human problem solving: searching for possible solutions, generating these possibilities out of other elements, and evaluating partial solutions and cues. From this standpoint there is nothing particularly novel about the theory. It rests its claims on other considerations:

1. It shows specifically and in detail how the processes that occur in human problem solving can be compounded out of elementary information processes, and hence how they can be carried out by mechanisms.

2. It shows that a program incorporating such processes, with appropriate organization, can in fact solve problems. This aspect of problem solving has been thought to be "mysterious" and unexplained because it was not understood how sequences of simple processes could account for the successful solution of complex problems. The theory dissolves the mystery by showing that nothing more need be added to the constitution of a successful problem solver.

Relation to Digital Computers

The ability to specify programs precisely, and to infer accurately the behavior they will produce, derives from the use of high-speed digital computers. Each specific theory—each program of

[1] Several studies of individual and group problem-solving behavior with logic problems have been carried out by O. K. Moore and Scarvia Anderson (5). The problems Moore and Anderson gave their subjects are somewhat different from those handled by our program, and hence a detailed comparison of behavior is not yet possible. We are now engaged, with Peter Houts, in replicating and extending the experiments of Moore and Anderson with human subjects and at the same time modifying our program to predict the human laboratory behavior in detail.

information processes that purports to describe some human behavior—is coded for a computer. That is, each primitive information process is coded to be a separate computer routine, and a "master" routine is written that allows these primitive processes to be assembled into any system we wish to specify. Once this has been done, we can find out exactly what behavior the purported theory predicts by having the computer "simulate" the system.

We wish to emphasize that we are not using the computer as a crude analogy to human behavior—we are not comparing computer structures with brains, nor electrical relays with synapses. Our position is that the appropriate way to describe a piece of problem-solving behavior is in terms of a program: a specification of what the organism will do under varying environmental circumstances in terms of certain elementary information processes it is capable of performing. This assertion has nothing to do—directly—with computers. Such programs could be written (now that we have discovered how to do it) if computers had never existed.[2] A program is no more, and no less, an analogy to the behavior of an organism than is a differential equation to the behavior of the electrical circuit it describes. Digital computers come into the picture only because they can, by appropriate programming, be induced to execute the same sequences of information processes that humans execute when they are solving problems. Hence, as we shall see, these programs describe both human and machine problem solving at the level of information processes.[3]

With this discussion of the relation of programs to machines and humans behind us, we can afford to relax into convenient, and even metaphoric, uses of language without much danger of misunderstanding. It is often convenient to talk about the behavior implied by a program as that of an existing physical mechanism doing things. This mode of expression is legitimate, for if we take the trouble to put any particular program in a computer, we have in fact a machine that behaves in the way prescribed by the program. Similarly, for concreteness, we will often talk as if our theory of problem solving consisted of statements about the ability of a computer to do certain things.

THE LOGIC THEORIST

We can now turn to an example of the theory. This is a program capable of solving problems in a particular domain—capable, specifically, of discovering proofs for theorems in elementary symbolic logic. We shall call this program the Logic Theorist (LT).[4] We assert that the behavior of this program, when the stimulus consists of the instruction that it prove a particular theorem, can be used to predict the be-

[2] We can, in fact, find a number of attempts in the psychological literature to explain behavior in terms of programs—or the prototypes thereof. One of the most interesting, because it comes relatively close to the modern conception of a computer program, is Adrian de Groot's analysis of problem solving by chess players (2). The theory of de Groot is based on the thought-psychology of Selz, a somewhat neglected successor to the Wurzburg school. Quite recently, and apparently independently, we find the same idea applied by Jerome S. Bruner and his associates to the theory of concept formation (1). Bruner uses the term "strategy," derived from economics and game theory, for what we have called a program.

[3] For a fuller discussion of this point see (9).

[4] In fact, matters are a little more complicated, for in the body of this paper we will consider both the basic program of LT and a number of variants on this program. We will refer to all of these variants, interchangeably, as "LT." This will not be confusing, since the exact content of the program we are considering at any particular point will always be clear from the context.

havior of (certain) humans when they are faced with the same problem in symbolic logic.

The program of LT was not fashioned directly as a theory of human behavior; it was constructed in order to get a program that would prove theorems in logic. To be sure, in constructing it the authors were guided by a firm belief that a practicable program could be constructed only if it used many of the processes that humans use. The fact remains that the program was not devised by fitting it directly to human data. As a result, there are many details of LT that we would not expect to correspond to human behavior. For example, no particular care was exercised in choosing the primitive information processes to correspond, point by point, with elementary human processes. All that was required in writing the program was that the primitive processes constitute a sufficient set and a convenient set for the type of program under study.

Since LT has been described in detail elsewhere (6, 8), the description will not be repeated here. It will also be unnecessary to describe in detail the system of symbolic logic that is used by LT. For those readers who are not familiar with symbolic logic, we may remark that problems in the sentential calculus are at about the same level of difficulty and have somewhat the same "flavor" as problems in high school geometry.[5]

Design of the Experiments

First we will describe the overt behavior of LT when it is presented with problems in elementary symbolic logic. In order to be concrete, we will refer to an experiment conducted on a digital computer. We take an ordinary general-purpose digital computer,[6] and store in its memory a program for interpreting the specifications of LT. Then we load the program that specifies LT. The reader may think of this program as a collection of techniques that LT has acquired for discovering proofs. These techniques range from the ability to read and write expressions in symbolic logic to general schemes for how a proof might be found.

Once we have loaded this program and pushed the start button, the computer, to all intents and purposes, *is* LT. It already knows how to do symbolic logic, in the sense that the basic rules of operation of the mathematics are already in the program (analogously to a human's knowing that "equals added to equals give equals" in elementary algebra).

We are now ready to give LT a task. We give it a list of the expressions (axioms and previously proved theorems) that it may take as "given" for the task at hand. These are stored in LT's memory. Finally, we present LT with another expression and instruct it to discover a proof for this expression.

From this point, the computer is on its own. The program plus the task uniquely determines its behavior. It attempts to find a proof—that is, it tries various techniques, and if they don't

[5] LT employs the sentential calculus as set forth in Chapters 1 and 2 of A. N. Whitehead and Bertrand Russell, *Principia Mathematica* (10)—the "classic" of modern symbolic logic. A simple introduction to the system of *Principia* will be found in (3).

[6] The experiments described here were carried out with the RAND JOHNNIAC computer. The JOHNNIAC is an automatic digital computer of the Princeton type. It has a word length of 40 bits, with two instructions in each word. Its fast storage consists of 4,096 words of magnetic cores, and its secondary storage consists of 9,216 words on magnetic drums. Its speed is about 15,000 operations per second. The programming techniques used are described more fully in (6). The experiments are reported in more detail in (7).

work, it tries other techniques. If LT finds a legitimate proof, it prints this out on a long strip of paper. There is, of course, no guarantee that it will find a proof; after working for some time, the machine will give up—that is, it will stop looking for a proof.

Now the experimenters know exactly what is in the memory of LT when it starts—indeed, they created the program. This, however, is quite different from saying that the experimenters can predict everything LT will do. In principle this is possible; but in fact the program is so complex that the only way to make detailed predictions is to employ a human to simulate the program by hand. (A human can do anything a digital computer can do, although it may take him considerably longer.)

1. As the initial experiment, we stored the axioms of *Principia Mathematica*, together with the program, in the memory of LT, and then presented to LT the first 52 theorems in Chapter 2 of *Principia* in the sequence in which they appear there. LT's program specified that as a theorem was proved it was stored in memory and was available, along with the axioms, as material for the construction of proofs of subsequent theorems. With this program and this order of presentation of problems, LT succeeded in proving 38 (73%) of the 52 theorems. About half of the proofs were accomplished in less than a minute each; most of the remainder took from one to five minutes. A few theorems were proved in times ranging from 15 minutes to 45 minutes. There was a strong relation between the times and the lengths of the proofs—the time increasing sharply (perhaps exponentially) with each additional proof step.

2. The initial conditions were now restored by removing from LT's memory the theorems it had proved. (Translate: "A new subject was obtained who knew how to solve problems in logic but was unfamiliar with the particular problems to be used in the experiment.") When one of the later theorems of Chapter 2 (Theorem 2.12) was presented to LT, it was not able to find a proof, although when it had held the prior theorems in memory, it had found one in about ten seconds.

3. Next, an experiment was performed intermediate between the first two. The axioms and Theorem 2.03 were stored in memory, but not the other theorems prior to Theorem 2.12, and LT was again given the task of proving the latter. Now, using Theorem 2.03 as one of its resources, LT succeeded—in fifteen minutes—where it had failed in the second experiment. The proof required three steps. In the first experiment, with all prior theorems available, the proof required only one step.

Outcome of the Experiments

From these three series of experiments we obtain several important pieces of evidence that the program of LT is qualitatively like that of a human faced with the same task. The first, and most important, evidence is that LT does in fact succeed in finding proofs for a large number of theorems.

Let us make this point quite clear. Since LT can actually discover proofs for theorems, its program incorporates a *sufficient* set of elementary processes arranged in a sufficiently effective strategy to produce this result. Since no other program has ever been specified for handling successfully these kinds of problem-solving tasks, no definite alternative hypothesis is available. We are well aware of the standard argument that "similarity of function does not imply similarity of process." However useful a caution this may be, it should not blind us to the fact that specification of a set of mechanisms sufficient to produce observed behavior

is strong confirmatory evidence for the theory embodying these mechanisms, especially when it is contrasted with theories that cannot establish their sufficiency.

The only alternative problem-solving mechanisms that have been completely specified for these kinds of tasks are simple algorithms that carry out exhaustive searches of all possibilities, substituting "brute force" for the selective search of LT. Even with the speeds available to digital computers, the principal algorithm we have devised as an alternative to LT would require times of the order of hundreds or even thousands of years to prove theorems that LT proves in a few minutes. LT's success does not depend on the "brute force" use of a computer's speed, but on the use of heuristic processes like those employed by humans.[7] This can be seen directly from examination of the program, but it also shows up repeatedly in all the other behavior exhibited by LT.

The second important fact that emerges from the experiments is that LT's success depends in a very sensitive way upon the order in which problems are presented to it. When the sequence is arranged so that before any particular problem is reached some potentially helpful intermediate results have already been obtained, then the task is easy. It can be made progressively harder by skipping more and more of these intermediate stepping-stones. Moreover, by providing a single "hint," as in the third experiment (that is, "Here is a theorem that might help"), we can induce LT to solve a problem it had previously found insoluble. All of these results are easily reproduced in the laboratory with humans. To compare LT's behavior with that of a human subject, we would first have to train the latter in symbolic logic (this is equivalent to reading the program into LT), but without using the specific theorems of Chapter 2 of *Principia Mathematica* that are to serve as problem material. We would then present problems to the human subject in the same sequence as to LT. For each new sequence we would need naive subjects, since it is difficult to induce a human subject to forget completely theorems he has once learned.

[7] A quantitative analysis of the power of the heuristics incorporated in LT will be found in (7).

PERFORMANCE PROCESSES IN THE LOGIC THEORIST

We can learn more about LT's approximation to human problem solving by instructing it to print out some of its intermediate results—to work its problems on paper, so to speak. The data thus obtained can be compared with data obtained from a human subject who is asked to use scratch paper as he works a problem, or to think aloud.[8] Specifically, the computer can be instructed to print out a record of the subproblems it works on and the methods it applies, successfully and unsuccessfully, while seeking a solution. We can obtain this information at any level of detail we wish, and make a correspondingly detailed study of LT's processes.

To understand the additional information provided by this "thinking aloud" procedure, we need to describe a little

[8] Evidence obtained from a subject who thinks aloud is sometimes compared with evidence obtained by asking the subject to theorize introspectively about his own thought processes. This is misleading. Thinking aloud is just as truly behavior as is circling the correct answer on a paper-and-pencil test. What we infer from it about other *processes* going on inside the subject (or the machine) is, of course, another question. In the case of the machine, the problem is simpler than in the case of the human, for we can determine exactly the correspondence between the internal processes and what the machine prints out.

more fully how LT goes about solving problems. This description has two parts: (*a*) specifying what constitutes a proof in symbolic logic; (*b*) describing the methods that LT uses in finding proofs.

Nature of a Proof

A proof in symbolic logic (and in other branches of logic and mathematics) is a sequence of statements such that each statement: (*a*) follows from one or more of the others that precede it in the sequence, or (*b*) is an axiom or previously proved theorem.[9] Here "follows" means "follows by the rules of logic."

LT is given four rules of inference:

Substitution. In a true expression (for example, "[p or p] implies p") there may be substituted for any variable a new variable or expression, provided that the substitution is made throughout the original expression. Thus, by substituting p or q for p in the expression "(p or p) implies p," we get: "([p or q] or [p or q]) implies (p or q)" but *not:* "([p or q] or p) implies p."

Replacement. In a true expression a connective ("implies," etc.) may be replaced by its definition in terms of other connectives. Thus "A implies B" is defined to be "not-A or B"; hence the two forms can be used interchangeably.

Detachment. If "A" is a true expression and "A implies B" is a true expression, then B may be written down as a true expression.

Syllogism (Chaining). It is possible to show by two successive applications of detachment that the following is also legitimate: If "a implies b" is a true expression and "b implies c" is a true expression, then "a implies c" is also a true expression.

Proof Methods

The task of LT is to construct a proof sequence deriving a problem expression from the axioms and the previously proved theorems by the rules of inference listed above. But the rules of inference, like the rules of any mathematical system or any game, are permissive, not mandatory. That is, they state what sequences *may* legitimately be constructed, not what particular sequence should be constructed in order to achieve a particular result (i.e., to prove a particular problem expression). The set of "legal" sequences is exceeding large, and to try to find a suitable sequence by trial and error alone would almost always use up the available time or memory before it would exhaust the set of legal sequences.[10]

To discover proofs, LT uses *methods* which are particular combinations of information processes that result in coordinated activity aimed at progress in a particular direction. LT has four methods (it could have more): *substi-*

[9] The axioms of symbolic logic and the theories that follow from them are all tautologies, true by virtue of the definitions of their terms. It is their tautological character that gives laws of logic their validity, independent of empirical evidence, as rules of inductive inference. Hence the very simple axioms that we shall use as examples here will have an appearance of redundancy, if not triviality. For example, the first axiom of *Principia* states, in effect, that "if any particular sentence (call it p) is true, or if that same sentence (p) is true, then that sentence (p) is, indeed, true"—for example, "if frogs are fish, or if frogs are fish, then frogs are fish." The "if---then" is trivially and tautologically true irrespective of whether p is true, for in truth frogs are not fish. Since our interest here is in problem solving, not in logic, the reader can regard LT's task as one of manipulating symbols to produce desired expressions, and he can ignore the material interpretations of these symbols.

[10] See (7). The situation here is like that in chess or checkers where the player knows what moves are legal but has to find in a reasonable time a move that is also "suitable"—that is, conducive to winning the game.

tution, detachment, forward chaining, and *backward chaining.* Each method focuses on a single possibility for achieving a link in a proof.

The substitution method attempts to prove an expression by generating it from a known theorem employing substitutions of variables and replacements of connectives.

The detachment method tries to work backward, utilizing the rule of detachment to obtain a new expression whose proof implies the proof of the desired expression. This possibility arises from the fact that if B is to be proved, and we already know a theorem of the form "A implies B," then proof of A is tantamount to proof of B.

Both chaining methods try to work backward to new problems, using the rule of syllogism, analogously to the detachment method. Forward chaining uses the fact that if "a implies c" is desired and "a implies b" is already known, then it is sufficient to prove "b implies c." Backward chaining runs the argument the other way: desiring "a implies c" and knowing "b implies c" yields "a implies b" as a new problem.

The methods are the major organizations of processes in LT, but they are not all of it. There is an executive process that coordinates the use of the methods, and selects the subproblems and theorems upon which the methods operate. The executive process also applies any learning processes that are to be applied. Also, all the methods utilize common subprocesses in carrying out their activity. The two most important subprocesses are the *matching* process, which endeavors to make two given subexpressions identical, and the *similarity test,* which determines (on the basis of certain computed descriptions) whether two expressions are "similar" in a certain sense (for details, cf. 8).

LT can be instructed to list its attempts, successful and unsuccessful, to use these methods, and can list the new subproblems generated at each stage by these attempts. We can make this concrete by an example:

Suppose that the problem is to prove "p implies p." The statement "(p or p) implies p" is an axiom; and "p implies (p or p)" is a theorem that has already been proved and stored in the theorem memory. Following its program, LT first tries to prove "p implies p" by the substitution method, but fails because it can find no similar theorem in which to make substitutions.

Next, it tries the detachment method. Letting B stand for "p implies p," several theorems are found of the form "A implies B." For example, by substitution of not-p for q, "p implies (q or p)" becomes "p implies (not-p or p)"; this becomes, in turn, by replacement of "or" by "implies": "p implies (p implies p)." Discovery of this theorem creates a new subproblem: "Prove A" —that is, "prove p." This subproblem, of course, leads nowhere, since p is not a universally true theorem, hence cannot be proved.

At a later stage in its search LT tries the chaining method. Chaining forward, it finds the theorem "p implies (p or p)" and is then faced with the new problem of proving that "(p or p) implies p." This it is able to do by the substitution method, when it discovers the corresponding axiom.

All of these steps, successful and unsuccessful, in its proof—and the ones we have omitted from our exposition, as well—can be printed out to provide us with a complete record of how LT executed its program in solving this particular problem.

SOME CHARACTERISTICS OF THE PROBLEM-SOLVING PROCESS

Using as our data the information provided by LT as to the methods it

tries, the sequence of these methods, and the theorems employed, we can ask whether its procedure shows any resemblance to the human problem-solving process as it has been described in psychological literature. We find that there are, indeed, many such resemblances, which we summarize under the following headings: set, insight, concept formation, and structure of the problem-subproblem hierachy.

Set

The term "set," sometimes defined as "a readiness to make a specified response to a specified stimulus" (**4**, p. 65), covers a variety of psychological phenomena. We should not be surprised to find that more than one aspect of LT's behavior exhibits "set," nor that these several evidences of set correspond to quite different underlying processes.

1. Suppose that after the program has been loaded in LT, the axioms and a sequence of problem expressions are placed in its memory. Before LT undertakes to prove the first problem expression, it goes through the list of axioms and computes a description of each for subsequent use in the "similarity" tests. For this reason, the proof of the first theorem takes an extra interval of time amounting, in fact, to about twenty seconds. Functionally and phenomenologically, this computation process and interval represent a *preparatory set* in the sense in which that term is used in reaction-time experiments. It turns out in LT that this preparatory set saves about one third of the computing time that would otherwise be required in later stages of the program.

2. *Directional set* is also evident in LT's behavior. When it is attempting a particular subproblem, LT tries first to solve it by the substitution method. If this proves fruitless, and only then, it tries the detachment method, then chaining forward, then chaining backward. Now when it searches for theorems suitable for the substitution method, it will not notice theorems that might later be suitable for detachment (different similarity tests being applied in the two cases). It attends single-mindedly to possible candidates for substitution until the theorem list has been exhausted; then it turns to the detachment method.

3. Hints and the change in behavior they induce have been mentioned earlier. Variants of LT exist in which the order of methods attempted by LT, and the choice of units in describing expressions, depend upon appropriate hints from the experimenter.

4. Effects from directional set occur in certain learning situations—as illustrated, for example, by the classical experiments of Luchins. Although LT at the present time has only a few learning mechanisms, these will produce strong effects of directional set if problems are presented to LT in appropriate sequences. For example, it required about 45 minutes to prove Theorem 2.48 in the first experiment because LT, provided with all the prior theorems, explored so many blind alleys. Given only the axioms and Theorem 2.16, LT proved Theorem 2.48 in about 15 minutes because it now considered a quite different set of possibilities.

The instances of set observable in the present program of LT are natural and unintended by-products of a program constructed to solve problems in an efficient way. In fact, it is difficult to see how we could have avoided such effects. In its simplest aspect, the problem-solving process is a search for a solution in a very large space of possible solutions. The possible solutions must be examined in *some* particular sequence, and if they are, then certain possible solutions will be examined before others. The par-

ticular rule that induces the order of search induces thereby a definite set in the ordinary psychological meaning of that term.

Preparatory set also arises from the need for processing efficiency. If certain information is needed each time a possible solution or group of solutions is to be examined, it may be useful to compute this information, once and for all, at the beginning of the problem-solving process, and to store it instead of recomputing it each time.

The examples cited show that set can arise in almost every aspect of the problem-solving process. It can govern the sequence in which alternatives are examined (the "method" set), it can select the concepts that are used in classifying perceptions (the "viewing" set), and it can consist in preparatory processes (the description of axioms).

None of the examples of set in LT relate to the way in which information is stored in memory. However, one would certainly expect such set to exist, and certain psychological phenomena bear this out—the set in association experiments, and so-called "incubation" processes. LT as it now stands is inadequate in this respect.

Insight

In the psychological literature, "insight" has two principal connotations: (*a*) "suddenness" of discovery, and (*b*) grasp of the "structure" of the problem, as evidenced by absence of trial and error. It has often been pointed out that there is no necessary connection between the absence of overt trial-and-error behavior and grasp of the problem structure, for trial and error may be perceptual or ideational, and no obvious cues may be present in behavior to show that it is going on.

In LT an observer's assessment of how much trial and error there is will depend on how much of the record of its problem-solving processes the computer prints out. Moreover, the amount of trial and error going on "inside" varies within very wide limits, depending on small changes in the program.

The performance of LT throws some light on the classical debate between proponents of trial-and-error learning and proponents of "insight," and shows that this controversy, as it is usually phrased, rests on ambiguity and confusion. LT searches for solutions to the problems that are presented it. This search must be carried out in some sequence, and LT's success in actually finding solutions for rather difficult problems rests on the fact that the sequences it uses are not chosen casually but do, in fact, depend on problem "structure."

To keep matters simple, let us consider just one of the methods LT uses —proof by substitution. The number of valid proofs (of *some* theorem) that the machine can construct by substitution of new expressions for the variables in the axioms is limited only by its patience in generating expressions. Suppose now that LT is presented with a problem expression to be proved by substitution. The crudest trial-and-error procedure we can imagine is for the machine to generate substitutions in a predetermined sequence that is independent of the expression to be proved, and to compare each of the resulting expressions with the problem expression, stopping when a pair are identical (cf. 7).

Suppose, now, that the generator of substitutions is constructed so that it is *not* independent of the problem expression—so that it tries substitutions in different sequences depending on the nature of the latter. Then, if the dependence is an appropriate one, the amount of search required on the average can be reduced. A simple strategy of this sort would be to try in the

axioms only substitutions involving variables that actually appear in the problem expression.

The actual generator employed by LT is more efficient (and hence more "insightful" by the usual criteria) than this. In fact, it works backward from the problem expression, and takes into account necessary conditions that a substitution must satisfy if it is to work. For example, suppose we are substituting in the axiom "p implies (q or p)," and are seeking to prove "r implies (r or r)." Working backward, it is clear that *if* the latter expression can be obtained from the former by substitution at all, then the variable that must be substituted for p is r. This can be seen by examining the first variable in each expression, without considering the rest of the expression at all (cf. 7).

Trial and error is reduced to still smaller proportions by the method for searching the list of theorems. Only those theorems are extracted from the list for attempted substitution which are "similar" in a defined sense to the problem expression. This means, in practice, that substitution is attempted in only about ten per cent of the theorems. Thus a trial-and-error search of the theorem list to find theorems similar to the problem expression is substituted for a trial-and-error series of attempted substitutions in each of the theorems.

In these examples, the concept of proceeding in a "meaningful" fashion is entirely clear and explicit. Trial-and-error attempts take place in some "space" of possible solutions. To approach a problem "meaningfully" is to have a strategy that either permits the search to be limited to a smaller subspace, or generates elements of the space in an order that makes probable the discovery of one of the solutions early in the process.

We have already listed some of the most important elements in the program of LT for reducing search to tolerable proportions. These are: (*a*) the description programs to select theorems that are "likely" candidates for substitution attempts; (*b*) the process of working backwards, which uses information about the goal to rule out large numbers of attempts without actually trying them. In addition to these, the executive routine may select the sequence of subproblems to be worked on in an order that takes up "simple" subproblems first.

Concepts

Most of the psychological research on concepts has focused on the processes of their formation. The current version of LT is mainly a performance program, and hence shows no concept formation. There is in the program, however, a clearcut example of the use of concepts in problem solving. This is the routine for describing theorems and searching for theorems "similar" to the problem expression or some part of it in order to attempt substitutions, detachments, or chainings. All theorems having the same description exemplify a common concept. We have, for example, the concept of an expression that has a single variable, one argument place on its left side, and two argument places on its right side: "p implies (p or p)" is an expression exemplifying this concept; so is "q implies (q implies q)."

The basis for these concepts is purely pragmatic. Two expressions having the same description "look alike" in some undefined sense; hence, if we are seeking to prove one of them as a theorem, while the other is an axiom or theorem already proved, the latter is likely construction material for the proof of the former.

Hierarchies of Processes

Another characteristic of the behavior of LT that resembles human problem-solving behavior is the hierarchical structure of its processes. Two kinds of hierarchies exist, and these will be described in the next two paragraphs.

In solving a problem, LT breaks it down into component problems. First of all, it makes three successive attempts: a proof by substitution, a proof by detachment, or a proof by chaining. In attempting to prove a theorem by any of these methods, it divides its task into two parts: first, finding likely raw materials in the form of axioms or theorems previously proved; second, using these materials in matching. To find theorems similar to the problem expression, the first step is to compute a description of the problem expression; the second step is to search the list of theorems for expressions with the same description. The description-computing program divides, in turn, into a program for computing the number of levels in the expression, a program for computing the number of distinct variables, and a program for computing the number of argument places.

LT has a second kind of hierarchy in the generation of new expressions to be proved. Both the detachment and chaining methods do not give proofs directly but, instead, provide new alternative expressions to prove. LT keeps a list of these subproblems, and, since they are of the same type as the original problem, it can apply all its problem-solving methods to them. These methods, of course, yield yet other subproblems, and in this way a large network of problems is developed during the course of proving a given logic expression. The importance of this type of hierarchy is that it is not fixed in advance, but grows in response to the problem-solving process itself, and shows some of the flexibility and transferability that seem to characterize human higher mental processes.

The problem-subproblem hierarchy in LT's program is quite comparable with the hierarchies that have been discovered by students of human problem-solving processes, and particularly by de Groot in his detailed studies of the thought methods of chess players (2, pp. 78–83, 105–111). Our earlier discussion of insight shows how the program structure permits an efficient combination of trial-and-error search with systematic use of experience and cues in the total problem-solving process.

SUMMARY OF THE EVIDENCE

We have now reviewed the principal evidence that LT solves problems in a manner closely resembling that exhibited by humans in dealing with the same problems. First, and perhaps most important, it is in fact capable of finding proofs for theorems—hence incorporates a system of processes that is sufficient for a problem-solving mechanism. Second, its ability to solve a particular problem depends on the sequence in which problems are presented to it in much the same way that a human subject's behavior depends on this sequence. Third, its behavior exhibits both preparatory and directional set. Fourth, it exhibits insight both in the sense of vicarious trial and error leading to "sudden" problem solution, and in the sense of employing heuristics to keep the total amount of trial and error within reasonable bounds. Fifth, it employs simple concepts to classify the expressions with which it deals. Sixth, its program exhibits a complex organized hierarchy of problems and subproblems.

COMPARISON WITH OTHER THEORIES

We have proposed a theory of the higher mental processes, and have shown how LT, which is a particular exemplar

of the theory, provides an explanation for the processes used by humans to solve problems in symbolic logic. What is the relation of this explanation to others that have been advanced?

Associationism

The broad class of theories usually labelled "associationist" share a generally behaviorist viewpoint and a commitment to reducing mental functions to elementary, mechanistic neural events. We agree with the associationists that the higher mental processes can be performed by mechanisms—indeed, we have exhibited a specific set of mechanisms capable of performing some of them.

We have avoided, however, specifying these mechanisms in neurological or pseudo-neurological terms. Problem solving—at the information-processing level at which we have described it—has nothing specifically "neural" about it, but can be performed by a wide class of mechanisms, including both human brains and digital computers. We do not believe that this functional equivalence between brains and computers implies any structural equivalence at a more minute anatomical level (e.g., equivalence of neurons with circuits). Discovering what neural mechanisms realize these information-processing functions in the human brain is a task for another level of theory construction. Our theory is a theory of the information processes involved in problem solving, and not a theory of neural or electronic mechanisms for information processing.

The picture of the central nervous system to which our theory leads is a picture of a more complex and active system than that contemplated by most associationists. The notions of "trace," "fixation," "excitation," and "inhibition" suggest a relatively passive electrochemical system (or, alternatively, a passive "switchboard"), acted upon by stimuli, altered by that action, and subsequently behaving in a modified manner when later stimuli impinge on it.

In contrast, we postulate an information-processing system with large storage capacity that holds, among other things, complex strategies (programs) that may be evoked by stimuli. The stimulus determines what strategy or strategies will be evoked; the content of these strategies is already largely determined by the previous experience of the system. The ability of the system to respond in complex and highly selective ways to relatively simple stimuli is a consequence of this storage of programs and this "active" response to stimuli. The phenomena of set and insight that we have already described and the hierarchical structure of the response system are all consequences of this "active" organization of the central processes.

The historical preference of behaviorists for a theory of the brain that pictured it as a passive photographic plate or switchboard, rather than as an active computer, is no doubt connected with the struggle against vitalism. The invention of the digital computer has acquainted the world with a device—obviously a mechanism—whose response to stimuli is clearly more complex and "active" than the response of more traditional switching networks. It has provided us with operational and unobjectionable interpretations of terms like "purpose," "set," and "insight." The real importance of the digital computer for the theory of higher mental processes lies not merely in allowing us to realize such processes "in the metal" and outside the brain, but in providing us with a much profounder idea than we have hitherto had of the characteristics a mechanism must possess if it is to carry out complex information-processing tasks.

Gestalt Theories

The theory we have presented resembles the associationist theories largely in its acceptance of the premise of mechanism, and in few other respects. It resembles much more closely some of the Gestalt theories of problem solving, and perhaps most closely the theories of "directed thinking" of Selz and de Groot. A brief overview of Selz's conceptions of problem solving, as expounded by de Groot, will make its relation to our theory clear.

1. Selz and his followers describe problem solving in terms of processes or "operations" (2, p. 42). These are clearly the counterparts of the basic processes in terms of which LT is specified.

2. These operations are organized in a strategy, in which the outcome of each step determines the next (2, p. 44). The strategy is the counterpart of the program of LT.

3. A problem takes the form of a "schematic anticipation." That is, it is posed in some such form as: Find an X that stands in the specified relation R to the given element E (2, pp. 44–46). The counterpart of this in LT is the problem: Find a *sequence of sentences* (X) that stands in the relation of *proof* (R) to the given *problem expression* (E). Similarly, the subproblems posed by LT can be described in terms of schematic anticipations: for example, "Find an expression that is 'similar' to the expression to be proved." Many other examples can be supplied of "schematic anticipations" in LT.

4. The method that is applied toward solving the problem is fully specified by the schematic anticipation. The counterpart in LT is that, upon receipt of the problem, the executive program for solving logic problems specifies the next processing step. Similarly, when a subproblem is posed—like "prove the theorem by substitution"—the response to this subproblem is the initiation of a corresponding program (here, the method of substitution).

5. Problem solving is said to involve (*a*) finding means of solution, and (*b*) applying them (2, pp. 47–53). A counterpart in LT is the division between the similarity routines, which find "likely" materials for a proof, and the matching routines, which try to use these materials. In applying means, there are needed both *ordering* processes (to assign priorities when more than one method is available) and *control* processes (to evaluate the application) (2, p. 50).

6. Long sequences of solution methods are coupled together. This coupling may be *cumulative* (the following step builds on the result of the preceding) or *subsidiary* (the previous step was unsuccessful, and a new attempt is now made) (2, p. 51). In LT the former is illustrated by a successful similarity comparison followed by an attempt at matching; the latter by the failure of the method of substitution, which is then followed by an attempt at detachment.

7. In cumulative coupling, we can distinguish *complementary* methods from *subordinated* methods (2, p. 52). The former are illustrated by successive substitutions and replacements in successive elements of a pair of logic expressions. The latter are illustrated by the role of matching as a subordinate process in the detachment method.

We could continue this list a good deal further. Our purpose is not to suggest that the theory of LT can or should be translated into the language of "directed thinking." On the contrary, the specification of the program for LT clarifies to a considerable extent notions whose meanings are only vague in the earlier literature. What the list illustrates is that the processes

that we observe in LT are basically the same as the processes that have been observed in human problem solving in other contexts.

PERFORMANCE AND LEARNING

LT is primarily a performance machine. That is to say, it solves problems rather than learning how to solve problems. However, although LT does not learn in all the ways that a human problem solver learns, there are a number of important learning processes in the program of LT. These serve to illustrate some, but not all, of the forms of human learning.

Learning in LT

By learning, we mean any more or less lasting change in the response of the system to successive presentations of the same stimulus. By this definition—which is the customary one—LT does learn.

1. When LT has proved a theorem, it stores this theorem in its memory. Henceforth, the theorem is available as material for the proof of subsequent theorems. Therefore, whether LT is able to prove a particular theorem depends, in general, on what theorems it has previously been asked to prove.

2. LT remembers, during the course of its attempt to prove a theorem, what subproblems it has already tried to solve. If the same subproblem is obtained twice in the course of the attempt at a proof, LT will remember and will not try to solve it a second time if it has failed a first.

3. In one variant, LT remembers what theorems have proved useful in the past in conjunction with particular methods and tries these theorems first when applying the method in question. Hence, although its total repertory of methods remains constant, it learns to apply particular methods in particular ways.

These are types of learning that would certainly be found also in human problem solvers. There are other kinds of human learning that are not yet represented in LT. We have already mentioned one—acquiring new methods for attacking problems. Another is modifying the descriptions used in searches for similar theorems, to increase the efficiency of those searches. The latter learning process may also be regarded as a process for concept formation. We have under way a number of activities directed toward incorporating new forms of learning into LT, but we will postpone a more detailed discussion of these until we can report concrete results.

What is Learned

The several kinds of learning now found in LT begin to cast light on the pedagogical problems of "what is learned?" including the problems of transfer of training. For example, if LT simply stored proofs of theorems as it found these, it would be able to prove a theorem a second time very rapidly, but its learning would not transfer at all to new theorems. The storage of *theorems* has much broader transfer value than the storage of *proofs*, since, as already noted, the proved theorems may be used as stepping stones to the proofs of new theorems. There is no mystery here in the fact that the transferability of what is learned is dependent in a very sensitive way upon the form in which it is learned and remembered. We hope to draw out the implications, psychological and pedagogical, of this finding in our subsequent research on learning.

CONCLUSION

We should like, in conclusion, only to draw attention to the broader implications of this approach to the study of information-processing systems. The heart of the approach is describing the

behavior of a system by a well specified program, defined in terms of elementary information processes. In this approach, a specific program plays the role that is played in classical systems of applied mathematics by a specific system of differential equations.

Once the program has been specified, we proceed exactly as we do with traditional mathematical systems. We attempt to deduce general properties of the system from the program (the equations); we compare the behavior predicted from the program (from the equations) with actual behavior observed in experimental or field settings; we modify the program (the equations) when modification is required to fit the facts.

The promise of this approach is several-fold. First, the digital computer provides us with a device capable of realizing programs, and hence, of actually determining what behavior is implied by a program under various environmental conditions. Second, a program is a very concrete specification of the processes, and permits us to see whether the processes we postulate are realizable, and whether they are sufficient to produce the phenomena. The vaguenesses that have plagued the theory of higher mental processes and other parts of psychology disappear when the phenomena are described as programs.

In the present paper we have illustrated this approach by beginning the construction of a thoroughly operational theory of human problem solving. There is every reason to believe that it will prove equally fruitful in application to the theories of learning, of perception, and of concept formation.

REFERENCES

1. BRUNER, J. S., GOODNOW, J., & AUSTIN, G. *A study of thinking.* New York: Wiley, 1956.
2. DE GROOT, A. *Het Denken van den Schaker.* Amsterdam: Noord-Hollandsche Uitgevers Maatschappij, 1946.
3. HILBERT, D., & ACKERMANN, W. *Principles of mathematical logic.* New York: Chelsea, 1950.
4. JOHNSON, D. M. *The psychology of thought and judgment.* New York: Harper, 1955.
5. MOORE, O. K., & ANDERSON, S. B. Search behavior in individual and group problem solving. *Amer. sociol. Rev.*, 1955, **19**, 702–714.
6. NEWELL, A., & SHAW, J. C. Programming the logic theory machine. *Proceedings Western Joint Computer Conference* (Institute of Radio Engineers), 1957, 230–240.
7. NEWELL, A., SHAW, J. C., & SIMON, H. A. Empirical explorations with the logic theory machine. *Proceedings Western Joint Computer Conference* (Institute of Radio Engineers), 1957, 218–230.
8. NEWELL, A., & SIMON, H. A. The logic theory machine: A complex information processing system. *Transactions on information theory* (Institute of Radio Engineers), 1956, Vol. IT-2, No. 3, 61–79.
9. SIMON, H. A., & NEWELL, A. Models, their uses and limitations. In L. D. White (Ed.), *The state of the social sciences.* Chicago: Univer. Chicago Press, 1956. Pp. 66–83.
10. WHITEHEAD, A. N., & RUSSELL, B. *Principia mathematica.* Vol. I. (2nd ed.) Cambridge: Cambridge Univer. Press, 1925.

(Received July 10, 1957)

Psychological Review
Vol. 65, No. 3, 1958

AVOIDANCE CONDITIONING THROUGH IRRADIATION: A NOTE ON PHYSIOLOGICAL MECHANISMS AND PSYCHOLOGICAL IMPLICATIONS

JACK ARBIT [1]

U. S. Army Leadership Human Research Unit, Fort Ord, California [2]

Recent experiments by Garcia et al. (11, 13, 14) appear to be of great interest to psychologists studying that aspect of learning called avoidance behavior. These studies have shown that exposure to Co^{60} acts as an unconditioned stimulus (US) to produce avoidance behavior to the cues associated with the presentation of the radiation. In a footnote to one of these reports it was noted that "the physiological mechanisms through which radiation is capable of acting as an unconditioned stimulus are as yet unknown" (13, p. 40). It is the purpose of this note to indicate possible physiological mechanisms through which this effect occurs and to comment upon the importance of this phenomenon for psychology.

Wynne and Solomon (22) have shown that interfering with the normal functioning of the peripheral autonomic nervous system (ANS) by surgical and pharmacological techniques interferes with an animal's ability to acquire an avoidance response which, when learned, in some cases undergoes spontaneous extinction. Auld (3) reports that tetraethylammonium, a drug which blocks impulses through the autonomic ganglia, has a depressive effect upon speed of response in an escape-avoidance situation, although the possibility of a confounding effect due to a depressant action of the drug upon motor activity has been noted by Brady (4). By using a serial discrimination problem and utilizing such measures as errors, trials to a criterion, and trials to extinction, one may evaluate the effects of autonomic block that are independent of the effects of the drug upon activity or motor capacity of the organism. Under these conditions it was found that tetraethylammonium as well as hexamethonium, a second drug which has as one of its major effects the block of impulses through the autonomic ganglia, interferes not only with learning, utilizing escape from shock as motivation (2), but also interferes to some extent with learning in less stressful situations (1). From these studies one may conclude, although tentatively because of a shortage of confirmatory evidence, that the cues associated with the normal functioning of the ANS are important in goal directed, integrated behavior. Whether this is true because of an increased drive state produced by ANS functioning (for a recent summary see 21) or because of the feedback to the cortex from the ANS (5) is of little import for the considerations presented in this paper.

In a thorough review of the relationships between irradiation injury and host defense mechanisms, Donaldson and Marcus (7) have shown that irradiation-induced stress produces an increase in adrenal cortical function. This conclusion is based upon the work of Dougherty and White (8), Patt et al. (20), North and Nims (19), and Hochman and Bloch-Frankenthal (15), who

[1] The author would like to thank Ernest Furchtgott and John Garcia for their comments on an early draft of this note.

[2] This paper was written while the author was on active duty with the U. S. Army. The opinions and conclusions contained in this report are those of the author and should in no way be construed as reflecting the views or endorsement of the Department of the Army. Present address: V. A. Research Hospital, Chicago.

found decreases in ascorbic acid after irradiation. Nims and Sutton (18) present data which indicate that the decrease in adrenal cholesterol level as well as the polydipsia which accompanies whole body radiation is the result of increased activity of the pituitary-adrenal system. Lasser and Stenstrom (16), working with humans following pelvic area irradiation, concluded that the adrenal cortex underwent definite changes in the course of irradiation, although these changes did not appear to be correlated with clinical "radiation sickness." [3]

If, then, the feedback associated with ANS functioning induced by irradiation has the same properties as the autonomic feedback associated with electric shock, a sudden loud noise, or any number of other stimuli, we would hypothesize that, just as the pairing of environmental cues with electric shock causes these cues to acquire fear-inducing properties, so the pairing of environmental cues with irradiation should have this effect. It is these cues, then, which, when later presented without electric shock or irradiation, produce avoidance behavior.

Gastric changes, general malaise, and nausea are also known to accompany irradiation (Furchtgott; [4] Davis et al., 6; Leary, 17) and may well cause ANS feedback as well as feedback through the proprioceptive fibers of the somatic nervous system. This feedback also may become associated with environmental cues in such a way that upon presentation of the environmental cues an avoidance response to these cues is manifest.[5]

The impact, from the psychologist's point of view, of the studies by Garcia et al. is twofold. First, it appears that a US may produce avoidance conditioning even when the subject is not aware directly of its action. Awareness in this sense refers to the functioning of particular peripheral somatic afferents. When using electric shock to the feet of a rat through a grid upon which it stands, pain receptors in the feet supply awareness of the location of the trauma. When the US is radiation, particularly at minimal dosages, the animal has no awareness of the trauma through the somatic nervous system. Some knowledge, however, that it is undergoing a stress may be mediated through the action of the ANS, the changes in gastric functioning, and the general malaise and nausea produced by irradiation.

Second, the use of radiation as a US provides a solution to the problem noted by Solomon and Brush (21) of the general lack in experimental psychology of aversive stimuli other than electric shock. In this area the implications of the studies by Garcia et al. for further research are vast indeed. In a T maze will rats learn to turn away from a goal box in which they have been irradiated? What is the function of the curve relating drive strength and dosage level? What are the temporal parameters associated with the use of radiation as a US? Will a rat learn an instrumental response, such as bar pressing, to turn off a radiation source? [6] A review of

[3] Due to differences in dosage and exposure time, these studies are not directly applicable to the work of Garcia et al. However, it will be assumed tentatively that the time-dose parameter used by Garcia et al. results in ANS effects similar to those produced in those studies concerned more explicitly with irradiation and ANS changes.

[4] Personal communication from E. Furchtgott, University of Tennessee, May, 1957.

[5] In a recent study, Garcia and Kimeldorf (10) also reach the conclusion that the physiological mechanism underlying avoidance conditioning through irradiation may be stimulation of the ANS, although their evidence is somewhat different from that presented in this paper.

[6] Since writing this paper, it has come to the author's attention that a program of research at the U. S. Naval Radiological Defense Laboratory has been undertaken with just such questions in mind (10, 12).

the behavioral effects of ionizing radiation may be found in Furchtgott (9). The relative importance of radiation as an aversive stimulus as compared to, let us say, bright light or intense sound, appears to be in the greater potentialities it offers for elucidating the role of the underlying physiological mechanisms in learning and the problem of learning without "awareness."

REFERENCES

1. ARBIT, J. Effects of drive strength and chemical block of autonomic impulses upon maze learning and retention. *Psychol. Rep.*, 1957, **3**, 91–94.
2. ARBIT, J. Shock motivated serial discrimination learning and the chemical block of autonomic impulses. *J. comp. physiol. Psychol.*, 1958, **51**, 199–201.
3. AULD, F., JR. The effects of tetraethylammonium on a habit motivated by fear. *J. comp. physiol. Psychol.*, 1951, **44**, 565–574.
4. BRADY, J. V. Does tetraethylammonium reduce fear? *J. comp. physiol. Psychol.*, 1953, **46**, 307–310.
5. DARROW, C. W. Psychological and psychophysiological significance of the electroencephalogram. *Psychol. Rev.*, 1947, **54**, 156–168.
6. DAVIS, R. T., MCDOWELL, A. A., DETER, C. W., & STEELE, J. P. Performance of rhesus monkeys on selected laboratory tasks presented before and after a large single dose of whole-body X radiation. *J. comp. physiol. Psychol.*, 1956, **49**, 20–26.
7. DONALDSON, D. M., & MARCUS, S. Aspects of the relationship between irradiation injury and mammalian host defense mechanism: A review. *Air Univer., Sch. Aviation Med.*, 1956, No. 56–50.
8. DOUGHERTY, T. F., & WHITE, A. Pituitary-adrenal cortical control of lymphocyte structure and function as revealed by experimental X-radiation. *Endocrinology*, 1946, **39**, 370–385.
9. FURCHTGOTT, E. Behavioral effects of ionizing radiations. *Psychol. Bull.*, 1956, **53**, 321–334.
10. GARCIA, J., & KIMELDORF, D. J. Temporal relationship within the conditioning of a saccharin aversion through radiation exposure. *J. comp. physiol. Psychol.*, 1957, **50**, 180–183.
11. GARCIA, J., KIMELDORF, D. J., & HUNT, E. L. Conditioned responses to manipulative procedures resulting from exposure to gamma radiation. *Radiation Res.*, 1956, **5**, 79–87.
12. GARCIA, J., KIMELDORF, D. J., & HUNT, E. L. Spatial avoidance behavior in the rat as a result of exposure to ionizing radiation. *Brit. J. Radiol.*, 1957, **30**, 318–321.
13. GARCIA, J., KIMELDORF, D. J., HUNT, E. L., & DAVIES, B. P. Food and water consumption of rats during exposure to gamma radiation. *Radiation Res.*, 1956, **4**, 33–41.
14. GARCIA, J., KIMELDORF, D. J., & KOELLING, R. A. Conditioned aversion to saccharin resulting from exposure to gamma radiation. *Science*, 1955, **122**, 157–158.
15. HOCHMAN, A., & BLOCH-FRANKENTHAL, L. The effect of low and high X-ray dosage on the ascorbic acid content of the suprarenal. *Brit. J. Radiol.*, 1953, **26**, 599–600.
16. LASSER, E. C., & STENSTROM, K. W. Radiation sickness: A study of its relation to adrenal cortical function and the absolute eosinophil count. *Amer. J. Roentgenol.*, 1954, **72**, 474–487.
17. LEARY, R. W. Food-preference changes of monkeys subjected to low-level irradiation. *J. comp. physiol. Psychol.*, 1955, **48**, 343–346.
18. NIMS, L. F., & SUTTON, E. Adrenal cholesterol, liver glycogen and water consumption of fasting and X-irradiated rats. *Amer. J. Physiol.*, 1954, **177**, 51–54.
19. NORTH, N., & NIMS, L. F. Time-dose study of biochemical responses of rats to X-radiation. *Federation Proc.*, 1949, **8**, 119–120.
20. PATT, H. M., SWIFT, M. N., TYREE, E. B., & JOHN, E. S. Adrenal response to total body X-radiation. *Amer. J. Physiol.*, 1947, **150**, 480–487.
21. SOLOMON, R. L., & BRUSH, E. S. Experimentally derived conceptions of anxiety and aversion. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Lincoln: Univer. Nebraska Press, 1956.
22. WYNNE, L. C., & SOLOMON, R. L. Traumatic avoidance learning: Acquisition and extinction in dogs deprived of normal peripheral autonomic function. *Genetic Psychol. Monogr.*, 1955, **52**, 241–284.

(Received May 31, 1957)

Psychological Review
Vol. 65, No. 3, 1958

THE PROBLEM OF REDUCTIONISM IN PSYCHOLOGY

RICHARD JESSOR [1]

University of Colorado

The recent upsurge of interest in physiological determinants and physical models of behavior has raised or reopened certain fundamental questions about the logical status of psychology as an autonomous discipline among the sciences. Some of the discussions may fairly be summarized as implying that physiological or physical concepts are in some sense more basic than those of psychology and that, therefore, causal explanation of behavior will utimately be expressed in those terms. The controversy over the nature of the hypothetical construct (HC) in psychological explanation is a current example. Krech (16) has insisted that HCs must have neurophysiological reference or locus, and that psychologists should be content to define such things as needs, tensions, or cognitive structures as neural events. While this particular view has been criticized (2, 15), and there is much agreement that neurophysiological content is irrelevant in HCs (20), the more general orientation—that fundamental explanation will be reductive—continues to be influential.

Psychologists are not alone in being concerned about this problem. Anthropologists such as Kroeber and White have felt called upon to defend the autonomy of their discipline against those who have tried to explain culture in terms of the concepts of psychology or even biology. (White caustically comments that some anthropologists "have sold their culturological birthright for a mess of psychiatric pottage" [25, p. xix]). In turn, prominent biologists such as Needham, Haldane, Woodger, and Bertalanffy have argued that their own discipline cannot be considered simply an application of chemistry and physics.

Obviously, an inherent attraction of reductive explanation is its implications for possible ways of unifying the separate scientific disciplines. Since the unity of science is an ultimate aim of many scientific workers, the reductive point of view is not likely to be abandoned in the absence of reasonable alternative approaches to that goal. A commitment to reductionism, either because of its implications for the unity of science, or because of the belief that it represents more fundamental explanation, undoubtedly influences the strategy of work of many scientists.[2] For this reason alone it would seem worth while to assess the doctrine in some detail.

Some considerations of the reductionism problem have tended to dismiss it. One basis for dismissal is the assertion that the answer to the problem is entirely empirical in nature, depending on the course of future developments of science about which speculation is admittedly dangerous. This view seems unwarranted; while the ultimate relations among the sciences will be an empirical outcome, at any given point in time, the relations among the sciences

[1] The author was a Social Science Research Council postdoctoral fellow at the University of California during the time in which this paper was prepared.

[2] Sound empirical work, of course, requires no defense, whether motivated by reductionistic aims or not. Nothing said in this paper should be interpreted as depreciating the value of empirical or theoretical efforts to bridge the gap between disciplines.

are legitimate and important questions for logical analysis. Another basis for dismissal, especially where the disciplines involved are psychology and physiology, is the adoption of the reductionistic view as a logical or an in-principle certainty. The soundness of this view, also, is open to question, as will be shown. The purpose of this paper, then, is to examine the problem of reductionism and to try to make a small beginning in separating issues which are logical in character from those which are empirical. In the course of our analysis, it will be contended that there are certain *logical* barriers to any present-day physiological reductionism.

THE DOCTRINE OF REDUCTIONISM

The essence of reductionism would seem to include four related general propositions. (*a*) The several disciplines or sciences may be considered as hierarchically ordered from, e.g., physics at the base through chemistry, biology, and psychology, to the social and historical disciplines at the top.[3] (*b*) The second essential aspect of reductionism is the proposition that the terms or concepts and the relations or laws of one discipline may fully and without loss of meaning be translated into or deduced from those of another discipline. (*c*) Such deduction or derivability proceeds only in one direction, from lower to higher [4] levels in the hierarchical ordering, and hence the term "reductionism"; terms and laws of the higher discipline are "reduced" to those of a lower one. Thus, in our earlier example, the psychological term "cognitive structure" is considered translatable into—deducible from—terms belonging to neurophysiology. (*d*) The final aspect is the implicit or explicit proposition that the lower the level of terms employed to explain a given phenomenon, the more causal or fundamental or basic the explanation. This is really only a corollary of the first point if certain assumptions about the nature of the hierarchical ordering are made. These four propositions together would seem to constitute the essential meaning of reductionism as a general doctrine. An adherent of that point of view may not, of course, subscribe to all of its aspects.

The primary focus of this paper is upon the issues attending the reduction of psychology to physiology, and our evaluation of the doctrine will for the most part be oriented toward that specific context. Within that context, the following comments are illustrative of the position which supports the doctrine. "Logically and in principle, physiological reduction is a certainty.

[3] Finer discriminations can, of course, easily be made by including the well-known "border" disciplines such as biochemistry or social psychology. But the fact that there is no sharp break between the sciences, and that it is frequently difficult to tell where one leaves off and the other begins, need not, in itself, challenge the autonomous existence of the several disciplines. What may be implied by ordering them in a hierarchy from lower to higher will be discussed shortly; for the moment it is only important to consider that this is one of the notions essential to the doctrine of reductionism.

[4] Terms in this paper referring to "position" in the hierarchy of the sciences, e.g., higher-lower, upward-downward, top-bottom, are by no means to imply any valuative judgment. The *meaning* of position in the hierarchy has been variously specified, for example, as referring to levels of abstraction or levels of integration, or as referring to the order of historical evolution within the universe of the subject matter of the sciences, or even to the order of historical emergence of the sciences themselves (**25**). As pointed out later, the concept of levels of science is not an analytically clear one (cf. **17**). For present purposes it is sufficient to take note of the existence in scientific discourse of such a hierarchical concept, and to recognize the traditional general ordering which places the physical sciences at the base, the biological sciences in the middle, and the social sciences at the top of the hierarchy.

Every bit of behavior and everything that can, like conscious contents, be defined in terms of behavior has its physiological correlate" (2, p. 442); "Relative to the 'molar' (or macro-) account given by behavioristic psychology, the neurophysiological account is a micro-description of the very same events and processes" (10, p. 623); and finally, ". . . molar behavioristics is in theory completely reducible to underlying neurophysiological principles. . . . A completely deterministic neurophysiology must of necessity permit derivation of all molar behavioral laws" (20, pp. 261–262). An obvious corollary of these statements is the logical reducibility of physiological principles in turn to those of chemistry and ultimately of physics.

Any challenge to these kinds of statements and their implications can be seen to require both logical and empirical arguments. We shall try to show where each kind is appropriate.

To begin with, to speak of reducing one discipline to another requires that the terms or concepts of the one be distinguishable from those of the other. This is not an easy requirement. Woodger (26) calls it an interesting methodological question to inquire how one knows what belongs to the language of neurology and what belongs to that of psychology. Similarly, in commenting on the new terms which will accrue as science develops, Hempel (12) notes that it is by no means certain that each of these terms will be readily classifiable as physical or nonphysical. What seems to be a necessary preliminary for our examination of physiological reductionism is some general criterion for separating or identifying the terms of psychology and physiology. If this is accomplished by defining psychology in a certain way, it should make apparent some logical barriers to the possibility of a physiological reduction.

A FUNCTIONAL DEFINITION OF PSYCHOLOGY

Admittedly, the definition of any scientific discipline is somewhat arbitrary; despite this, it is certainly possible to obtain adequate agreement on criteria for segregating one discipline from another or for grossly circumscribing the domain of a particular science. Among psychologists there is considerable agreement that the scope and subject matter of concern is the behavior of whole, human[5] organisms. The difficulty with such a general statement is that the term "behavior" is not without ambiguity; psychologists have been notoriously neglectful in providing a systematic definition of a response. This laxness has, it is felt, obscured the conceptual boundary between psychology and physiology. The present discussion of behavior as a psychological concept follows the implicit orientation of all functional behavior theories, and more specifically, the approaches of Kantor (14), Brunswik (5, 6, 7), and Tolman (24).

The central point of these approaches is that behavior, *qua* psychological, refers to an organism-environment *interaction* or *relationship*. Tolman (24) specifically states that the complete identification of any behavior-act requires reference to its relation to particular goal-objects and the intervening means-objects with which it has commerce. Kantor's interbehaviorism makes essentially the same point. In considering the question of where to establish the boundaries of a behavior, Bentley (1) similarly concludes that behavior must be recognized as a transdermal process or event whose description must immediately and functionally include the environmental and situational settings.

[5] The writer assumes that the study of animal behavior by psychologists is merely propaedeutic to a science of human behavior.

Within this framework, then, behavior viewed psychologically is interactional or relational in nature; its specification or identification at the referential level requires the specification of a particular context and a set of relationships thereto. Our definition of psychology, therefore, excludes the study of organisms or physical environments per se, and behavior may not be referred to either alone. The laws of behavior of a discipline so defined refer to the dynamics of organism-environment functional interaction. The terms or concepts of those laws describe what may be called an interaction or behavior space.

This definition of psychology was undertaken as a means of providing a criterion for deciding whether a behavior term, or a law involving such a term, properly belongs to psychology or physiology. Our criterion requires of psychological terms that they have immediate reference to a functionally defined environment or context. Before drawing the implications of these considerations for reductionism, a further comment on this kind of definition may be in order. Some objection may be raised to the relational or transdermal character of the definition in that there is provided no palpable locus for a psychologically defined behavior. Those who raise such a query seem to be operating within what Woodger (27) picturesquely describes as a "finger and thumb" philosophy of metaphysics, i.e., the notion that a thing is real or exists only if it can in principle be picked up between the finger and thumb. Interactions or relations, though not simple physical objects, are nevertheless real and concrete and can be precisely specified by the conditions and course of their occurrence.

INCOMPLETE DERIVABILITY OF TERMS

It may now be asserted that the reference of psychology, as defined, is profoundly different from that of physiology. The terms and laws of the latter refer to intraorganismic or intradermal processes, or, at most, relations between an organism or its parts and the space defined by physics. They refer, in short, to the functioning of anatomical structures or processes described by body parameters. To state one of the major contentions of this paper, the absence in contemporary physiology of any systematic terms for describing the functional environment or context of behavior would seem to preclude, *on logical grounds alone,* any complete reduction of psychology to physiology. These necessary and sufficient conditions for the terms of psychology cannot be described in physiological (or physical) terms alone. This "incompleteness" of the lower discipline's language, in being able to specify only the physiological correlates but not the environmental correlates of a behavior or response, constitutes the logical impediment. This point obviously requires elaboration and further support.

Let us take as an example an occurrence described in common-sense, nonsystematic language—the wave of an arm as two persons pass each other—and compare the systematic descriptions of it by the two disciplines being discussed. The systematic language of the physiologist enables him to rely on only body parameters or physical terms. He may thus speak of arm-displacement, changes in muscle-tension, metabolic rate, blood-volume distribution, and neural reactivity. In none of these terms is it systematically possible to take cognizance of the social context. On the other hand, the psychologist may describe the event as waving a greeting to a friend, or even, since micromediation is generally of little interest to him, and equifinality generally

taken for granted,[6] simply as greeting a friend. If psychological laws refer to interactions between organisms and functionally defined environments, only these latter descriptions of events can logically lead to the achievement of such laws. A discipline such as physiology, lacking such contextual terms, cannot therefore be considered *logically* equivalent; hence its adequacy cannot be guaranteed, even in principle, as a complete reduction base for psychology.

One of the strongest empirical sources of support for some of these considerations may be found in the research analyses of Campbell (8, 9) and Smedslund (22, 23) evoked by the recent controversy over "what is learned." As mentioned earlier, the term "response" has been conceptually neglected; this neglect seems, in turn, to be related to the inadequate attention given to the environment or context of behavior by many psychological theorists (cf. 13 in this connection). Brunswik (5, 6, 7), however, is one of those who has called attention to the problem in his emphasis upon *distal* achievement in the adjustment of an organism to its ecology. Beginning with this orientation, and appraising a variety of learning experiments, especially those dealing with transposition, both Campbell and Smedslund conclude that the learned response must be defined in environmental terms in order successfully to accommodate—predict—the actual research findings. Campbell states that ". . . the learned response is to be essentially defined in terms of a shift in the organism-environment relationship rather than a motor response defined in terms of organism or body parameters alone" (9, p. 105).

Scrutiny of the literature on reductionism has showed only meager if any attention to the specific issue raised in this paper. Two writers may be cited whose remarks, made in other contexts, are pertinent to the logical soundness of our contention. In his Tarner lectures (26), Woodger stresses the role of environment as a determinant at various biological levels from zygote to whole human organism. He rejects the body-mind dualism in favor of speaking about persons and developing a person-language. The notion of person requires, for its very definition, environmental specification, and the latter, he observes, requires words belonging to sociology. To treat persons otherwise, i.e., in terms of body parameters, loses sight of this fact: "But this is the only way in which we can treat them so long as we confine ourselves to the physical sciences, since these sciences do not provide a vocabulary for speaking about them in any other way" (26, p. 261). Hempel has discussed the problem in relation to the possible derivability of all the laws of empirical science from those of physics (a logical corollary of physiological reductionism). Affirming that not all the terms of empirical science are definable by means of the vocabulary of physics, he asserts that ". . . a law containing, say, certain biological terms cannot, in general, be logically derived from a set of physical laws, which contain no biological terms at all . . ." (12, pp. 320–321).[7] Both of these writers seem in support of our point of view that there are logical barriers to any thorough-going physiological reductionism. The barriers reside in the absence of terms in the "lower" discipline which would enable the logical derivability of descriptions of the functional context of behavior and, thereby, the derivation of the laws of psychology.

[6] That is, the same psychological event may be served by (partly constituted of) an almost infinite variety of different physiological events.

[7] To achieve such a derivation requires some law connecting the biological concepts with the physical concepts. "But those connecting laws are not purely physical in character" (12, p. 321). And they have the character of *empirical* laws.

At least two kinds of questions can be raised about our analysis and are worth consideration at this point. The first of these is the possibility of overcoming the logical problem we have raised simply by incorporating into physiology the category of terms we have suggested that it lacks. This is frequently what is implied by the phrase "in principle" in assertions about the possibility of reductionism. But as Sellars (21) has pointed out in another connection, this makes the entire problem an empty truism in that it involves a "tacit redefinition" of physiological theory to encompass psychology. Such a redefinition is a statement about the *future* state of the sciences involved, and it therefore transfers the discussion from logical to empirical grounds. That is, the question of whether a future physiology will be able to encompass psychology depends entirely upon the nature and direction of ongoing empirical development of *both* disciplines. It is to be noted here, too, that the meaning of reduction is always and only relative to a given state of the disciplines concerned. To suggest adding terms to physiology in order to make psychology deducible from it implies the "elevation" or "expansion" of physiology just as much as it implies reduction of psychology. (In connection with these issues, see the Meehl-Sellars discussion of the logic of emergentism, 19.) Finally, it seems unlikely that scientific theory develops or advances by simple accretion of the terms of other theories. We will return to this point shortly, in considering reductionism and the unity of science.

The second kind of question which may be raised is implicit in the earlier quote from Bergmann. He bases his affirmation of the logical certainty of physiological reductionism on the proposition that ". . . everything that can be defined in terms of behavior has its physiological correlate" (2, p. 442). The essential point of this position would be that the environmental reference of behavior upon which we based our definition must, if effective, be represented within the organism in its physiology or, especially, neurophysiology. This is, in a sense, a proximal approach to behavior rather than a distal one. Certain arguments may be brought to bear against this position, such as the abscence of strict one-to-one proximal-distal correlation—e.g., in perception—and the theoretical significance of vicarious functioning or equifinality. The major reply which may be made, however, is that the "recovery" of actual behavioral phenomena from physiological correlates requires the conceptual coordination of these correlates to environmental contexts. Thus the issue raised originally reappears, the necessity for terms to represent or describe the context of behavior. That such coordination can or will be accomplished in the future is an empirical rather than a logical problem, and therefore not a logical certainty.

Our discussion of reductionism up to this point has concerned itself largely with the logical problems inherent in the second and third essential aspects of the doctrine as outlined at the start of this paper. Proper attention to the first and fourth aspects would extend the paper beyond practical space limitations. Instead, we shall simply sketch some of the issue requiring attention.

The first aspect has to do with the hierarchical ordering of the several scientific disciplines. Despite the widespread acceptance of the hierarchy notion—witness the frequent reference to "levels" of science, the employment of terms like "basic" to contrast disciplines, and the characterization of certain disciplines as "emergent" from others—it is not an analytically clear concept. Kroeber (17) remarks on the absence of any adequate attempt to ex-

amine systematically what the levels constitute or mean in terms of a theory of knowledge. Most of the discussions of what is meant by levels resolve into two positions which are generally considered to exclude one another. One point of view conceives the levels of science to be a matter of methodology only, i.e., to refer to the kinds of procedures employed by the various sciences, the size of their units of analysis, etc. The other point of view considers the levels to refer to substantive differences in the events or phenomena dealt with by the various disciplines. Psychologists will be familiar with this contrast from the Littman-Rosen (18) analysis of the molar-molecular problem. A third possibility is that these two positions are correlated rather than mutually exclusive, namely, that substantively different events require particularly appropriate methodological procedures for useful analysis.

Each of these three views has certain implications for the doctrine of reductionism. For example, the methodological position would seem to favor the doctrine, since it assumes the events or phenomena to be the same and only the descriptions of them to be different. If only the descriptions differ, e.g., in size of unit, it should be logically possible to reduce the larger units to their smaller constituents. On the other hand, to assert a substantive difference between levels would seem to be unfavorable to reductionism. The events or phenomena of higher levels are considered different from—not the same ones as—those of the lower levels, and therein lies the difficulty in reducing descriptions of one kind of event to those of another kind. No vitalistic or dualistic considerations need be involved in speaking of events as different; certainly organic events may in general be separated from inorganic ones, for example. Feigl's double-language theory (10) of the mind-body problem espouses the methodological position in insisting that the factual reference of the mentalistic, behavioristic and neurophysiological languages or levels of description is identical, i.e., involves the very same events and processes. Others concerned with the problem seem to adopt the substantive position; thus White (25) speaks of culture as a distinct class of events, a distinct order of phenomena. Woodger tentatively concludes that ". . . perhaps, in spite of superficial appearances, person-acts and behaviour [defined in physical or physiological terms] are not quite the same things . . ." (26, p. 284). The position taken by the present writer in differentiating psychological from physiological terms may be seen as compatible with the substantive view. What may account for substantive differences, whether the substantive vs. methodological distinction itself is useful or defensible, the issues involved in emergentism, and any fuller analysis of the implications of the hierarchy of sciences for reductionism must be deferred for discussion elsewhere.

The final aspect of the reductionism doctrine has to do with the idea that causal explanation is advanced by the employment of terms of a lower-level discipline. In psychology, this notion identifies causal explanation with neurophysiological reference. The key reason for this approach would seem to be the belief in the higher levels as simply derivable from, or applications of, terms and laws of lower disciplines. Once this idea is abandoned, causal explanation could just as logically proceed upward to sociological and anthropological concepts. Knowledge of causality is probably best divorced from the hierarchy of sciences notion and considered instead to vary with the scope of the network in which any concept is embedded.

We have tried to examine some of the logical and empirical problems related

to the doctrine of physiological reductionism. Some of the considerations have led us to doubt that psychology can be reduced to physiology, though certainly many of the questions are of an empirical sort. Nothing thus far asserted in any way denies the possibility or the desirability of the unification of the sciences, the synthesis of psychology and physiology. What is questioned is that such synthesis must proceed by reduction, by one discipline devouring or incorporating the other. All of the sciences are developing, and their influence upon each other does not only proceed upward from physics. Bertalanffy (**3**, **4**), for example, sees certain biological developments, such as the notions of open systems, requiring extensions of the conceptual system of physics. Unity of science for him refers only to the structural isomorphy of laws in the different fields of science, the approach of General Systems Theory. Psychologists can also see the conceptual system of physiology as influenced by the data and concepts of psychology (**11**). Thus, unification of the sciences may proceed from above as well as from below. The continued autonomous development of each of the sciences will at least serve to specify the properties required of any synthetic unifying scheme. This is probably part of what Brunswik had in mind when he wrote: "Insistence on reduction as a universal goal of science can only result in blighted spots on the landmap of scientific enterprise" (**7**, p. 237).

SUMMARY

This paper has had as its aim the instigation of renewed attention to the doctrine of reductionism, especially in terms of its implications for the relationship of physiology and psychology. Despite the empirical character of the ultimate answer, it is asserted that the questions involved in the doctrine may properly be the concern of a logical analysis. After briefly sketching four propositions which constitute the essential notions of reductionism, the argument focused upon the logical possibility of a complete translatability or derivability of the concepts and laws of psychology from those of physiology. The central contention was that the latter, lacking terms to describe the behavioral environment, was logically inadequate as a base for a thoroughgoing reduction of the former. The remainder of the paper commented upon the relationship of the doctrine to the idea of a hierarchical ordering of the sciences and to the possibility of achieving a unification of science.

REFERENCES

1. BENTLEY, A. F. The behavioral superfice. *Psychol. Rev.*, 1941, **48**, 39–59.
2. BERGMANN, G. Theoretical psychology. In C. P. Stone (Ed.), *Annu. Rev. Psychol.*, 1953, **4**, 435–458.
3. BERTALANFFY, L. VON. Problems of general systems theory. *Human Biol.*, 1951, **23**, 302–312.
4. BERTALANFFY, L. VON. *Problems of life.* New York: Wiley, 1952.
5. BRUNSWIK, E. The conceptual focus of some psychological systems. *J. unif. Sci.*, 1939, **8**, 36–49.
6. BRUNSWIK, E. The conceptual framework of psychology. Chicago: Univer. Chicago Press, 1952. (*Int. Encycl. unif. Sci.*, Vol. 1, No. 10.)
7. BRUNSWIK, E. In defense of probabilistic functionalism: A reply. *Psychol. Rev.*, 1955, **62**, 236–242.
8. CAMPBELL, D. T. Operational delineation of "what is learned" via the transposition experiment. *Psychol. Rev.*, 1954, **61**, 167–174.
9. CAMPBELL, D. T. Adaptive behavior from random response. *Behavioral Sci.*, 1956, **1**, 105–110.
10. FEIGL, H. The mind-body problem in the development of logical empiricism. In H. Feigl & May Brodbeck (Eds.), *Readings in the philosophy of science.* New York: Appleton-Century-Crofts, 1953. Pp. 612–626.
11. HEBB, D. O. *The organization of behavior.* New York: Wiley, 1949.

12. HEMPEL, C. G. General systems theory and the unity of science. *Human Biol.*, 1951, **23**, 313–322.
13. JESSOR, R. Phenomenological personality theories and the data language of psychology. *Psychol. Rev.*, 1956, **63**, 173–180.
14. KANTOR, J. R. Preface to interbehavioral psychology. *Psychol. Rec.*, 1942, **5**, 173–193.
15. KESSEN, W., & KIMBLE, G. A. "Dynamic systems" and theory construction. *Psychol. Rev.*, 1952, **59**, 263–267.
16. KRECH, D. Dynamic systems, psychological fields, and hypothetical constructs. *Psychol. Rev.*, 1950, **57**, 283–290.
17. KROEBER, A. L. *The nature of culture.* Chicago: Univer. Chicago Press, 1952.
18. LITTMAN, R. A., & ROSEN, E. Molar and molecular. *Psychol. Rev.*, 1950, **57**, 58–65.
19. MEEHL, P. E., & SELLARS, W. The concept of emergence. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science.* Vol. I. Minneapolis: Univer. Minnesota Press, 1956. Pp. 239–252.
20. ROZEBOOM, W. W. Mediation variables in scientific theory. *Psychol. Rev.*, 1956, **63**, 249–264.
21. SELLARS, W. Empiricism and the philosophy of mind. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science.* Vol. I. Minneapolis: Univer. Minnesota Press, 1956. Pp. 253–329.
22. SMEDSLUND, J. The problem of "What is learned?" *Psychol. Rev.*, 1953, **60**, 157–158.
23. SMEDSLUND, J. *Multiple-probability learning.* Oslo: Akademisk Forlag, 1956.
24. TOLMAN, E. C. *Purposive behavior in animals and men.* Berkeley: Univer. California Press, 1949.
25. WHITE, L. A. *The science of culture.* New York: Farrar, Strauss, 1949.
26. WOODGER, J. H. *Biology and language.* Cambridge: Cambridge Univer. Press, 1952.
27. WOODGER, J. H. *Physics, psychology and medicine.* Cambridge: Cambridge Univer. Press, 1956.

(Received July 22, 1957)

Psychological Review
Vol. 65, No. 3, 1958

THE STATUS OF INHIBITION IN HULL'S SYSTEM: A THEORETICAL REVISION[1]

H. GWYNNE JONES

Institute of Psychiatry (Maudsley Hospital), University of London

Hull (**6, 7, 8**) clearly distinguishes between a habit ($_sH_R$) and a drive state (D) and postulates that these interact in a multiplicative manner so as to generate a reaction potential ($_sE_R$). Thus:

$$_sE_R = f(D \times {}_sH_R)$$

In his discussion of factors affecting response decrement, Hull further postulates the development of two forms of inhibition: an innate, fatigue-produced *reactive inhibition* (I_R) and *conditioned inhibition* ($_sI_R$).

The former is clearly conceived of and defined as a negative drive state, while the latter is considered a negative habit phenomenon. Despite this distinction, and in conflict with his earlier insistence on a multiplicative relationship between habits and drives, Hull postulates that these two variables interact in an additive fashion to produce inhibitory potential ($\dot{I}_R$). When such an inhibitory potential is operative it is subtracted from reaction potential to give a resultant or net reaction potential ($_s\bar{E}_R$). Thus:

$$_s\bar{E}_R = f[(D \times {}_sH_R) - (I_R + {}_sI_R)]$$

There has been considerable criticism of this formulation by such writers as Gleitman et al. (**5**) and Koch (**3**). Osgood (**10**) and Zeaman (**15**) similarly recognize the inconsistency of the equation and propose that $_sI_R$, as a negative habit, be subtracted directly from $_sH_R$, a positive habit, while continuing to allow I_R to be subtracted directly from $_sE_R$. Iwahara (**9**) recognizes that I_R, as a drive, must enter into some form of multiplicative relationship, but, in order to avoid certain implications of his argument, he then abandons $_sI_R$ as a habit phenomenon and proposes to consider it only as a secondary drive. Such an abandonment of the concept of a negative habit would involve a radical revision of much existing theory, while its retention is not imcompatible with the achievement of internal consistency within the Hullian system.

Iwahara bases his rejection upon only one possible modification of the basic equation, i.e., that inhibitory potential ($\dot{I}_R$) be considered as the product rather than the sum of reactive and conditioned inhibition. Symbolically:

$$\dot{I}_R = I_R \times {}_sI_R$$

He points out that this implies that the inhibitory potential would be zero when either I_R or $_sI_R$ is zero, a consequence clearly inconsistent with the known facts of learning. Conditioned inhibition, however, may, as a habit, also be energized by positive drives, e.g., when response decrement persists after long rest subsequent to massed practice. Similarly, reactive inhibition may interact with positive habits to lower reaction potential when no conditioned inhibition is present, as at the commencement of massed practice or extinction.

[1] This article is based upon a paper presented to the XV International Congress of Psychology at Brussels in August 1957. The proposed equation was earlier made use of by Eysenck (**4**) in a theoretical analysis of "warm-up" phenomena.

Thus it is suggested that net reaction potential may well be determined by the multiplicative interaction of a *resultant* drive state ($D - I_R$), and a *resultant* habit state ($_sH_R - {}_sI_R$), so that:

$$_s\bar{E}_R = f[(D - I_R) \times ({}_sH_R - {}_sI_R)]$$

The habit term in this equation is consistent with Osgood's and Zeaman's suggestion, while the drive term is consistent with Kimble's (**12**) experimentally validated proposition that the amount of reactive inhibition a subject will tolerate is a function of his drive level. This equation is mathematically equivalent to:

$$_s\bar{E}_R = f[(D \times {}_sH_R) - (I_R \times {}_sH_R) - (D \times {}_sI_R) + (I_R \times {}_sI_R)]$$

the fourth term of which represents a paradoxical addition to positive reaction potential from the interaction of two inhibitory variables. Similarly, if both resultant drive and habit states are negative, the resultant reaction potential would be positive. Such a state of affairs would occur rarely in practice, but may furnish a theoretical explanation of such phenomena as the "ultraparadoxical" phase of inhibition observed by Pavlov (**11**). The various inhibitory states of the organism induced by Pavlov in various ways may be considered as inhibitory drive states akin to reactive inhibition. In advanced stages of such inhibition, Pavlov noted that positive conditioned stimuli tended to lose their effect, whereas "well developed negative stimuli" (i.e., when $_sI_R$ was predominant) "acquired definite excitatory properties."

Any amendment of the Hullian equation must serve, not only to make Hullian theory internally consistent, but also to generate theorems consistent with empirical data. One such

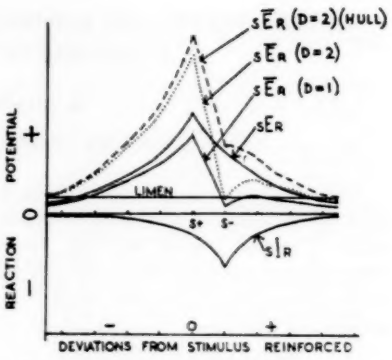

FIG. 1. Discrimination gradients at two levels of drive. See text for complete explanation.

theorem, derived from the present amendment, refers to stimulus discrimination learning, around which an apparently adequate theory in Hullian terms has been developed, particularly by Spence (**8, 14**).

The solid lines in Fig. 1 conventionally illustrate the algebraic summation theory of discrimination,[2] in which net reaction potential ($_s\bar{E}_R$) is a result of the subtraction of generalized inhibition ($_sI_R$) from generalized positive reaction potential ($_sE_R$). At the stage of learning illustrated, the net reaction potential attached to the nonreinforced or negative stimulus is well below the threshold for response, and the discrimination will be almost perfect.[3]

[2] The positive and negative generalization gradients in the figure were not calculated by formula but merely estimated, by eye, from textbook examples. Errors in reproduction and resulting differences between the two gradients will account for the unconventional hump in the curves of net reaction potentials. These are irrelevant to the argument in this paper.

[3] Hull claims that $_s\bar{E}_R$ cannot fall below $_sL_R$, the reaction threshold, because, when $_s\bar{E}_R = {}_sL_R$, no further responses occur and so no further $_sI_R$ develops. This argument neglects the effect of oscillation potential.

Let us now consider the effect of increasing drive in this situation; and, for convenience, let us assume that drive is, in fact, doubled. According to orthodox Hullian theory, this would have the effect of doubling the positive reaction potential but the conditioned inhibition would remain unchanged. The generalized net reaction potential (${}_s\bar{E}_R$) would then be represented by the broken line in Fig. 1. The net potential attached to the nonreinforced stimulus is in fact higher than at the commencement of discrimination learning in the low drive situation: the discrimination would have broken down completely. Even with lesser increases of drive, its multiplicative relationship with habit strength would insure severe impairment of discrimination.

If, however, the amended equation is employed, the doubled drive would have the effect of doubling both positive and negative potentials. The generalized net potential would then be represented by the dotted line in Fig. 1 and, at any point along the stimulus dimension, would be double that in the low drive conditions. At the nonreinforced stimulus value this would involve only a slight increment in potential, as the initial value is already subthreshold. Discrimination, while somewhat impaired, would not be expected to break down completely.

Empirical evidence is clearly in favor of the prediction from the amended equation. Skinner (**13**), for example, reports that under low hunger drive rats achieve almost perfect discrimination, performing bar-pressing responses when the box is dark and not pressing when the box is lighted. Under increased food deprivation the discriminations do not break down, but become impaired in that there is some "leaking" of positive responses to the negative stimulus.

Another situation in which the alternative equations would yield different predictions as to the effects of changes in drive is that in which habit strength is equal to conditioned inhibition. Calvin et al. (**1**) carried out an experiment to show that, despite continued reinforcement and under conditions of repeated massed practice, rats will ultimately refuse to traverse a runway for food and that this inhibition is retained over lengthy periods of time without such practice. Calvin's original experiment has been severely criticized by Dinsmoor (**2**) but, in a recent unpublished study, D. C. Kendrick achieved essentially similar results, using thirst motivation and controlling the factors relevant to Dinsmoor's criticisms.

If animals, trained in this fashion, are retested under conditions of increased drive, Hull's equation indicated that a reinstatement of the running performance should occur, whereas under decreased drive the inhibition should be more effective than in the original condition. According to the amended equation, change of drive in either direction, in that it affects ${}_sH_R$ and ${}_sI_R$ equally, should have no effect on the balance between these two variables. Any change of drive, in that it alters the stimulus situation, may affect performance, but running should be no more frequent with increased than with decreased drive. In a pilot study, employing ten animals and testing this prediction, Kendrick found that three animals responded positively under increased drive. Two of these and one other responded under decreased drive. No responses occurred when

When ${}_s\bar{E}_R = {}_sL_R$, owing to oscillation, the probability of response evocation will be 0.5. Only when ${}_s\bar{E}_R$ falls appreciably below ${}_sL_R$ will this probability fall to a low level.

drive level was unchanged. A more rigorous follow-up of this finding is now in progress.

SUMMARY

An amendment of the Hullian equation is proposed which, for both excitatory and inhibitory potentials, is consistent with Hull's distinction between habit and drive states and the postulated multiplicative interaction between the two. Theorems derived from this equation are compared with the analagous consequences of Hull's formulation in the light of available empirical data.

REFERENCES

1. CALVIN, A. D., CLIFFORD, L. T., CLIFFORD, B., BOLDEN, L., & HARVEY, J. Experimental validation of conditioned inhibition. *Psychol. Rep.*, 1956, **2**, 51–56.
2. DINSMOOR, J. D. Absurdum revisited: A comment. *Psychol. Rep.*, 1956, **2**, 255–256.
3. ESTES, W. K., KOCH, S., MACCORQUODALE, K., MEEHL, P. E., MUELLER, C. G., SCHOENFELD, W. N., & VERPLANCK, W. S. *Modern learning theory*. New York: Appleton-Century-Crofts, 1954.
4. EYSENCK, H. J. "Warm-up" in pursuit rotor learning as a function of the extinction of conditioned inhibition. *Acta psychol.*, 1956, **12**, 349–370.
5. GLEITMAN, H., NACHMIAS, J., & NEISSER, U. The *S-R* reinforcement theory of extinction. *Psychol. Rev.*, 1954, **61**, 23–33.
6. HULL, C. L. *The principles of behavior*. New York: Appleton-Century-Crofts, 1943.
7. HULL, C. L. *Essentials of behavior*. New Haven: Yale Univer. Press, 1951.
8. HULL, C. L. *A behavior system*. New Haven: Yale Univer. Press, 1952.
9. IWAHARA, S. Hull's concept of inhibition: A revision. *Psychol. Rep.*, 1957, **3**, 9–10.
10. OSGOOD, C. E. *Method and theory in experimental psychology*. New York: Oxford Univer. Press, 1953.
11. PAVLOV, I. P. *Conditioned reflexes*. London: Oxford Univer. Press, 1927.
12. KIMBLE, G. A., & BILODEAU, G. A. Work and rest as variables in cyclical motor learning. *J. exp. Psychol.*, 1950, **40**, 248–253.
13. SKINNER, B. F. *The behavior of organisms*. New York: Appleton-Century-Crofts, 1938.
14. SPENCE, K. W. The differential response in animals to stimuli varying within a single dimension. *Psychol. Rev.*, 1937, **44**, 430–444.
15. ZEAMAN, D. Response latency as a function of the amount of reinforcement. *J. exp. Psychol.*, 1949, **39**, 466–483.

(Received September 10, 1957)

THE SOURCES OF VALUE

By Stephen C. Pepper

A general hypothesis of the origins and functions of values in human life, drawing relevant data from psychology, anthropology, political science, economics, sociology, and other fields. Basic to the book is the concept of "selective systems," natural dynamic structures operating to eliminate errors and accumulate correct results. The author shows how such structures produce traditional value systems and operate in the behavior of both persons and groups. Of interest to psychologists, anthropologists, philosophers, and all concerned with how man meets the challenges of his world. *746 pages, $8.50*

At your bookstore

UNIVERSITY OF CALIFORNIA PRESS

Address: Berkeley 4, California

Person Perception and Interpersonal Behavior

Edited by Renato Tagiuri and Luigi Petrullo. The first book in English to deal exclusively with these matters. These contributions by a distinguished body of psychologists and anthropologists have systematically marked out areas of theory as well as of research. $7.50

Psychotherapy by Reciprocal Inhibition

Joseph Wolpe, M.D. The direct product of modern learning theory, this pioneer method of psychotherapy has achieved quick and lasting results. This book includes the theoretical, experimental, and clinical material that forms the background to this approach; the techniques of therapy; a statistical analysis of 210 patients. $5.00

Psychology in the Soviet Union

Edited by Brian Simon. "It acquaints English and American psychologists with some novel interpretations of psychological concepts, as well as with some original methods in attacking the problems of behavior."—*Science.* $6.00

Naven

Gregory Bateson. Second Edition. A brilliant new chapter in this study of the culture of a New Guinea tribe relates the ideas of the book to modern theories of psychiatry. $6.00

Stanford University Press
Stanford, California

The Child within the Group

AN EXPERIMENT IN SELF-GOVERNMENT

Marion E. Turner. With a Foreword by Theodore M. Newcomb. A record and analysis of a unique experiment that provides new information on the development of self-control among children, and conclusions on how individuals learn to be guided by group-made rules. "It took a bit of a miracle to produce this work—in fact, a double miracle. That a teacher should have made this experience possible was unusual enough; that the teacher should also have recorded it is almost too good to be true. Miss Turner has succeeded in teaching something important not only to her children but also to her readers."—From the Foreword. $3.00

A Classified Bibliography of Gerontology and Geriatrics

SUPPLEMENT ONE, 1949–1955

Nathan W. Shock. "A meticulous, well-organized, and exhaustive bibliography essential to anyone working in the field covered."—*Journal of Chronic Diseases.* $15.00

Trends in Gerontology

Nathan W. Shock. Second Edition. An accurate and comprehensive view of the varied problems of aging, including the psychological aspects of aging. $4.50

Stanford University Press
Stanford, California